HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

BACHELORARBEIT

Herr
Christian Schulz

Einsatz der Clusteranalyse für die Auswertung von Messungen mit dem Gleichgewichtskoordinationssystem Equilus alpha

2015

Fakultät **Angewandte Computer- und Biowissenschaften**

BACHELORARBEIT

Einsatz der Clusteranalyse für die Auswertung von Messungen mit dem Gleichgewichtskoordinationssystem Equilus alpha

Autor:
Christian Schulz

Studiengang:
Angewandte Mathematik

Seminargruppe:
MA12w1-B

Erstprüfer:
Herr Prof. Dr. Egbert Lindner

Zweitprüfer:
Herr Dr. Norman Bitterlich

Mittweida, September 2015

Bibliografische Angaben

Schulz, Christian: Einsatz der Clusteranalyse für die Auswertung von Messungen mit dem Gleichgewichtskoordinationssystem Equilus alpha, 50 Seiten, 1 Abbildung, Hochschule Mittweida, University of Applied Sciences, Fakultät Angewandte Computer- und Biowissenschaften

Bachelorarbeit, 2015

Referat

I. Inhaltsverzeichnis

II. Abbildungsverzeichnis

III. Tabellenverzeichnis

IV. Vorwort

In der heutigen Zeit ist es von großer Bedeutsamkeit den menschlichen Körper gesund und fit zu halten. Aus diesem und weiteren Gründen werden viele Diagnosegeräte und Verfahren entwickelt, um eine frühzeitige Erkennung der Krankheit und deren Therapie bereitzustellen.

Diese Arbeit beschäftigt sich mit dem Gleichgewichtskoordinationssystem Equilus alpha, welches mit verschieden Tests in der Lage ist, den Gleichgewichtssinn quantitativ zu analysieren. Die Clusteranalyse soll nun als Unterstützung dienen und bei der Bewertung der einzelnen Tests helfen.

Wie jede Arbeit wäre auch diese Bachelorarbeit nicht ohne die Unterstützung der Professoren der Fachhochschule Mittweida, insbesondere meinem Betreuer Herr Prof. Dr. Lindner, des Betriebes Medizin & Service, durch meinen Betreuer im Unternehmen Herr Dr. Bitterlich und durch meine Kommilitonen der Fachrichtung Mathematik an der Fachhochschule Mittweida entstanden. Ihnen allen möchte ich meinen besonderen Dank aussprechen.

Christian Schulz

1 Einleitung

Das Unternehmen Medizin & Service produziert das Gleichgewichtskoordinationssystem Equilus alpha. Mit diesem Produkt ist man in der Lage die Balance eines Menschen zu messen. Der Ursprung kommt aus der Physiotherapie, wo man die Behandlung von Strörungen im Gleichgewichtsgefühl beseitigen und somit den Menschen ein besseres Lebensgefühl vermitteln möchte.

Mit Hilfe dieser Arbeit soll nun der Einsatz der Clusteranalyse für die Auswertung der Tests des Gleichgewichtskoordinationssystems Equilus alpha vorgenommen werden. Zunächst werden dafür die beiden Tests Messung nach John und der Beugetest vorgestellt. Die Messung nach John ist eine Messung die bereits im Equilus-System integriert ist. Sie unterteilt sich in vier Einzelmessungen, wo die verschiedenen Subsysteme des menschlichen Körpers überprüft werden. Der Beugetest besteht auch bewussten Bewegung, um das Gleichgewichtsgefühl beim Beugen zu kontrollieren.

Auf Grundlage der gewonnenen Daten wird ein Clustering durchgeführt. Eine Einleitung bzw. Einführung erfolgt im Kapitel 3, welches die Grundlagen der Clusteranalyse näher erläutert. Im Anschluss werden konkrete Algorithmen für die einzelnen Probleme vorgestellt, nämlich die Ausreißererkennung der Daten und die Partitionierung der Daten.

Der Abschluss bildet einen Ausblick in die Bewertung von Clusterings und die Kenngrößen die eine quantitative Auswertung der Messung ermöglicht. Das letzte Kapitel soll einen Eindruck vermitteln inwiefern die Arbeit weitergeführt werden kann und eine Zusammenfassung der Arbeit darstellt.

2 Messungen mit dem Gleichgewichtskoordinationssystem Equilus alpha

Das Gleichgewichtskoordinationssystem Equilus alpha, im Folgenden Equilus genannt, ist eine Messeinheit, um den Gleichgewichtssinn bzw. Koordinationsstörungen bei Patienten zu bewerten. Sie dient als Unterstützung für den Arzt oder Therapeuten, welcher mit dem Gerät in der Lage ist, eine Form der Diagnostik durchzuführen. Dabei wird dem Patienten anhand der Messung verdeutlicht, ob Defizite vorliegen oder nicht.

2.1 Messprinzip

Für die Durchführung einer Messung stellt sich der Patient auf die Messplatte und muss dort für einen bestimmten Zeitraum in einem unbewegten Stand verweilen. Die Messplatte liegt auf einer weichen Unterlage, sodass durch die unbewussten Körperbewegungen eine „Verkippung“ erfolgt. Die Sensoren in der Messplatte registrieren diese Bewegung, indem sie die Winkel in äquidistanten Zeitabständen messen und diese in x-y-Koordinaten an den PC weiterleiten. Aufgrund dieser Eigenschaft ist das Gerät in der Lage, kleinste Verlagerungen des Körperschwerpunktes zu erfassen und zu visualisieren. Durch die Echtzeitverarbeitung der Daten ist der Arzt oder Therapeut während der Messung in der Lage, Auslenkungen zu beobachten und diese mit der momentanen Körperhaltung des Patienten zu vergleichen. Somit kann eine gezielte Diagnose durch einen Arzt bzw. Therapeuten stattfinden. Weil die ermittelten Daten in eine externe Datenbank gespeichert werden können, ist eine spätere und auch ortsunabhängige Auswertung möglich, um zum Beispiel die Fortschritte zu dokumentieren und auch eine Verlaufskurve der Person zu erstellen.

Ein weiterer Vorteil ist auch die Einsetzbarkeit des Equilus als Therapiegerät. Somit können spielerisch Defizite in der Koordination behoben werden oder Schwindelpatienten können selbst eine Verbesserung ihrer Wahrnehmung des Schwindelgefühls hervorrufen, um aktiv die unbeabsichtigte Bewegung kontrollieren zu lernen. Zusammengefasst kann der Patient lernen, den Gleichgewichtssinn und Bewegungssinn besser zu steuern. Stabilitätsübungen und ein Training der Reaktionen des menschlichen Körpers können vorgenommen werden, was eine Verbesserung des Balancegefühls herbeiführt. Durch eine Bewegungsanalyse kann eine Stabilisierung der verschiedenen Gelenke eintreten und des Weiteren haben die Übungen allgemein eine steigende Konzentrationsfähigkeit für den Patienten zur Folge.

2.2 Messung nach John

Die Messung nach John besteht aus vier Einzelmessungen. Während dieser Messungen sollte die Person eine ruhige Postion auf der Messplatte einnehmen. Dabei empfiehlt es sich die Knie nicht durch zu strecken, sondern leicht nach vorne zu beugen. Dies hilft der Person den Stand zu verbessern beziehungsweise kann somit ein Schwindelpatient schneller und sicherer wieder in die Ausgangsposition gelangen und den Schwindelanfall überwinden. Darüber hinaus befindet sich durch diese Haltung der Körperschwerpunkt etwas nach vorne verlagert, wodurch das Gewicht gleichmäßig auf dem Fuß verteilt ist. Wenn die Person aber mit durchgestreckten Beinen auf dem Equilus steht, werden nur die Fersen belastet. Diese Angaben zur Körperhaltung gelten in der Regel für alle Tests mit dem Equilus.

	Zustand der Augen	BalancePad	Dauer der Messung
erste Messung	geöffnet	ohne	30 Sekunden
zweite Messung	geschlossen	ohne	30 Sekunden
dritte Messung	geöffnet	mit	30 Sekunden
vierte Messung	geschlossen	mit	30 Sekunden

Tabelle 2.1: Messung nach John

Die in der aufgeführten Tabelle beschriebene Durchführung der Messungen, soll die einzelnen Subsysteme des Menschen beeinflussen und dadurch den Gleichgewichtssinn des Menschen charakterisieren. Für das Gleichgewicht des Menschen existieren drei Subsysteme im Körpers: das visuelle System, das somatosensorische System und das vestibuläre System. Unter dem visuellem System versteht man die Verknüpfung des Gleichgewichtsorgans mit den Augenmuskeln, welches dem Menschen ermöglicht ein stabiles Bild der Umgebung wahrzunehmen und dabei gleichzeitig den Kopf zu bewegen. Das vestibuläre System ist wichtig für die Wahrnehmung der Position, das Gleichgewicht und der Bewegung des Menschen im Raum. Dadurch wird eine Regulation der Augen möglich, was eine Orientierung und geordnete Körperhaltung des Menschen im Raum zur Folge hat. Das somatosensorische System oder auch Tiefensensibilität genannt, besitzt die Funktion der Wahrnehmung der Reize aus dem Körperinneren. Des Weiteren dient die Tiefensensibilität die Eigenwahrnehmung des Körpers im Raum festzustellen. Es unterteilt sich in Positionssinn, Kraftsinn und Bewegungssinn. Der Positionssinn gibt wieder Informationen über die Lage des Körpers, der Stellung der Gelenke und des Kopfes im Raum. Der Kraftsinn stellt die Informationen über Anspannung der Muskeln und Sehnen zur Verfügung. Der Bewegungssinn ermöglicht eine Erkennung der Bewegungsrichtung und der Bewegungsempfindung des Menschen.

Mittels der Messung nach John lassen sich diese Systeme untersuchen und sind in der folgenden Tabelle zusammengefasst dargestellt.

Durchgeführte Messung	visuell	somatosensorisch	vestibulär
ohne BalancePad, Augen geöffnet	•	•	•
ohne BalancePad, Augen geschlossen		•	•
mit BalancePad, Augen geöffnet	•		•
mit BalancePad, Augen geschlossen			•

Tabelle 2.2: abgeprüfte Subsysteme bei der Messung nach John

2.3 Beugetest

Das Ziel des Beugetests ist ein bewusste Veränderung des Körperschwerpunktes. Während des Tests steht die Person auf der Messplatte in zunächst normaler, ruhiger Haltung (ca. 10 Sekunden lang). In der zweiten Hälfte beugt sich die Person nach vorn, um den Körperschwerpunkt zu verlagern. Dabei soll sich die Person gerade nach vorne beugen und sollte nicht absichtlich nach links oder rechts abweichen. Diese nach vorn gebeugte Position ist bis zum Schluss einzuhalten.

In der zweiten Durchführung des Beugetests wird ein BalancePad zur Hilfe genommen. Der Ablauf setzt sich wie oben beschrieben zusammen. Zuerst eine ruhige, entspannte Position einnehmen und dann nach ca. 10 Sekunden sich nach vorne beugen. Diese Stellung muss die restliche Zeit gehalten werden.

Der Beugetest kann auch erweitert werden, bei der die Person im Anschluss der gebeugten Haltung sich wieder in die Grundhaltung zurück begibt. Dies bedeutet, dass man im ersten Drittel der Messung ruhig auf der Messplatte stehen soll, im zweiten Drittel sich nach vorne beugt und diese Position beibehalten muss und im letzten Drittel sich wieder in die Grundhaltung zurück bewegen soll.

Der Arzt oder Therapeut muss dabei beachten, dass bei den zwei Messungen (einmal mit und ohne BalancePad) jeweils zwei Punktewolken entstehen. Falls diese sich zu stark überlagern, muss der Test wiederholt werden. Wird trotzdem eine Analyse durchgeführt, könnten fehlerhafte Berechnungen entstehen und somit den Test verfälschen. Für den Test muss die Messplatte in einem bestimmten Abstand $dist$ von einer Wand aufgestellt werden, sodass der Patient eine optische Barriere hat und die Angst des Fallens minimiert wird.

$$dist = \lfloor \text{Körpergröße} \cdot (cos(70^\circ) - 0,15) - 5,5\,cm \rfloor$$

Der Abstand $dist$ setzt sich zusammen aus der Körpergröße, dem Fallwinkel α und einiger Abstände der Messplatte $(5,5\,cm)$. Die Körpergröße ist durch den Patienten gegeben. Der Fallwinkel $\alpha = \sphericalangle B'AC$ berechnet sich aus der Körpergröße, in der rechten Abbildung ist dies durch die Strecke $\overline{AB'}$ dargestellt, indem man die Höhe des Körperschwerpunkts ermittelt kann man über ein rechtwinkliges Dreieck den Winkel α ausrechnen. Dieser hat in der Regel eine Größe von $70°$. Die $0,15$ ist der Anteil der Körpergröße, welche im Allgemeinen der Fußlänge entspricht, was in der Abbildung mit der Strecke x dargestellt ist. In der folgenden Tabelle sind ein paar Abstände aufgeführt.

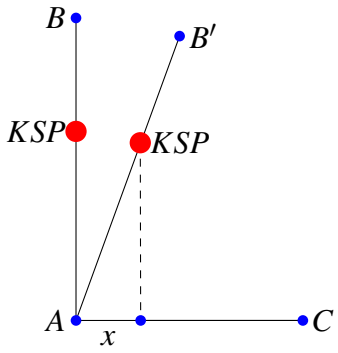


Abbildung 2.1: Darstellung der Berechnung des Maßes $dist$

Körpergröße $[cm]$	$dist$ $[cm]$
150	23
160	25
170	27
180	29
190	30
200	32

Tabelle 2.3: Abstand der Messplatte $(dist)$ in Abhängigkeit der Körpergröße

Der Patient sollte versuchen, sich bis zur Wand vorzubeugen. Der Beugetest bezieht sich nur auf die Verlagerung des Körperschwerpunktes. Es ist dabei nur zu beachten, dass die Person aus der Ausgangsposition in die Beugeposition gelangt und dabei ein genügend großer Abstand zwischen den Punktwolken entsteht.

3 Einführung in die Clusteranalyse

Das Ziel der Clusteranalyse ist es, aus einer gegeben Menge von Objekten oder Daten, eine Zuteilung in einzelne Cluster zu finden. Andere Bezeichnungen für das Wort Cluster wären Punktwolken, Klassen, Gruppen, Partitionen oder auch Teilmengen. Um eine Vorstellung der Problematik zu erhalten, wird im Folgenden die Anzahl der möglichen Gruppierungen von n Objekten auf k Cluster ermittelt. Der Begriff Partition bzw. Cluster wird verwendet, wie er in [Tit00] und [Mic09] definiert ist.

Definition 3.1 (Partition) Es sei M eine endliche Menge. Eine Menge $\{M_1, M_2, ..., M_k\}$ von Teilmengen von M ist eine Partition von M, wenn gilt:

1. $M_i \neq \emptyset$ für $i = 1, ..., k$,
2. $i \neq j \Longrightarrow M_i \cap M_j = \emptyset$ und
3. $\bigcup_{i=1}^{k} M_i = M$.

Ausformuliert bedeutet die Definition 3.1, dass eine Partition von M eine Zerlegung von M in k paarweise disjunkte, nichtleere Teilmengen ist. Diese werden, in der Kombinatorik, Blöcke einer Partition genannt.

Definition 3.2 (Stirling Zahl 2. Art) Die Stirling Zahl 2. Art $\left\{ {n \atop k} \right\}$ gibt die Anzahl der Partitionen einer n-elementigen Menge mit genau k Blöcken an.

Dabei gelten folgende Eigenschaften:

(1) $\left\{ {n \atop k} \right\} = 0$, falls $k > n$

(2) $\left\{ {n \atop 1} \right\} = \left\{ {n \atop n} \right\} = 1$

(3) $\left\{ {n \atop n-1} \right\} = \binom{n}{2}$

(4) $\left\{ {n \atop 2} \right\} = \frac{2^n - 2}{2} = 2^{n-1} - 1$

(5) $\left\{ {n \atop k} \right\} = \left\{ {n-1 \atop k-1} \right\} + k \left\{ {n-1 \atop k} \right\}$

Definition 3.3 (Bell Zahl) Die Bell Zahl B_n ist die Anzahl der Partitionen einer n-elementigen

Menge, die definiert ist als Summe der Stirling Zahl 2. Art.

$$B_n = \sum_{k=1}^{n} \left\{ {n \atop k} \right\}$$

Die Anzahl der Gruppierungen lassen sich mit Hilfe der Stirling Zahl zweiter Art ausrechnen. In der folgenden Tabelle werden für $n = 1, ..., 10$ und für $k = 1, ..., 10$ die Anzahl der Möglichkeiten aufgeführt.

$n\backslash k$	1	2	3	4	5	6	7	8	9	10	B_n
1	1	0	0	0	0	0	0	0	0	0	1
2	1	1									2
3	1	3	1								5
4	1	7	6	1							15
5	1	15	25	10	1						52
6	1	31	90	65	15	1					203
7	1	63	301	350	140	21	1				877
8	1	127	966	1701	1050	266	28	1			4140
9	1	255	3025	7770	6951	2646	462	36	1		21147
10	1	511	9330	34105	42525	22827	5880	750	45	1	115975

Tabelle 3.1: Auflistung der Stirling Zahlen zweiter Art

Aus der Tabelle 3.1 ist zu entnehmen, dass für $n = 10$ Objekten die in $k = 4$ Klassen bzw. Cluster aufgeteilt werden, es insgesamt 34105 Möglichkeiten gibt. Mit Hilfe der Bell Zahl lässt sich die Gesamtanzahl der Gruppierungen ausdrücken. Für die Anzahl von $n = 10$ Objekten existieren insgesamt 115975 Möglichkeiten diese zu gruppieren.

Für die Einteilung in Cluster müssen die beiden folgenden Bedingungen gelten.

Definition 3.4 (Homogenität aus [BPW10]) Die Homogenität innerhalb eines Clusters sagt aus, dass die Daten innerhalb eines gleichen Clusters möglichst ähnlich sein sollen.

Definition 3.5 (Heterogenität aus [BPW10]) Die Heterogenität zwischen den Clustern sagt aus, dass die Daten, die unterschiedlichen Clustern zugeordnet sind, möglichst verschieden sein sollen.

In diesem Zusammenhang stellt sich die Frage, wann die Daten innerhalb eines Clusters ähnlich sind und wann nicht. Die Antwort auf diese Frage kann nur mit Hilfe der Daten selbst geklärt werden. Der Begriff Ähnlichkeit ist dabei abhängig von den Daten. In der Regel sind die Daten reellwertige Vektoren und man kann verschiedene Abstände als Grundlage für die Abhängigkeit definieren, beschrieben in [GM98].

Definition 3.6 (Euklidischer Abstand) Der euklidische Abstand für zwei Vektoren $x, y \in \mathbb{R}^n$, ist durch die euklidische Norm $\|x-y\|_2$ bestimmt. Sei der Vektor $x = (x_1, ..., x_n)$ und der Vektor $y = (y_1, ..., y_n)$ gegeben, so gilt:

$$d(x,y) = \|x-y\|_2 = \sqrt{(x_1-y_1)^2 + \cdots + (x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}.$$

Definition 3.7 (Mahalanobis-Distanz) Die Mahalanobis-Distanz für zwei Vektoren $x, y \in \mathbb{R}^n$ und der Kovarianzmatrix S, ist definiert durch

$$d(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}.$$

Definition 3.8 (L_r-Distanz) Die L_r-Distanz für zwei Vektoren $x, y \in \mathbb{R}^n$ und $r \in \mathbb{N}$ ist definiert durch

$$d(x,y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r\right)^{\frac{1}{r}}.$$

Des Weiteren müssen die Abstandsmaße den Eigenschaften eines metrischen Raumes unterliegen.

Definition 3.9 (Metrischer Raum) Sei X eine beliebige Menge. Eine Abbildung $d : X \times X \to \mathbb{R}$ heißt Metrik auf X, wenn für beliebige Elemente $x, y, z \in X$ folgende Bedingungen erfüllen:

(1)	$d(x,y) \geq 0$ und $d(x,y) = 0 \Leftrightarrow x = y$	positive Definitheit
(2)	$d(x,y) = d(y,x)$	Symmetrie
(3)	$d(x,y) \leq d(x,z) + d(z,y)$	Dreiecksungleichung

(X,d) heißt metrischer Raum, wenn d ein Metrik auf X ist.

Die Bedingungen eines metrischen Raums werden durch den euklidischen Abstand erfüllt, was durch einfaches Nachrechnen zu zeigen ist. Des Weiteren wird $X = \mathbb{R}^n$ festgelegt.
Beweis:
zu (1):
Für $x \neq y$ gilt

$$d(x,y) = \|x-y\|_2 = \sqrt{\sum_{i=1}^{n} \underbrace{(x_i-y_i)^2}_{\geq 0}} > 0. \tag{3.1}$$

Für $x = y$ gilt

$$d(x,y) = d(x,x) = \|x-x\|_2 = \|\vec{0}\|_2 = \sqrt{\sum_{i=1}^{n} 0^2} = \sqrt{0} = 0. \tag{3.2}$$

zu (2):

$$d(x,y) = \|x-y\|_2 = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} = \sqrt{\sum_{i=1}^{n}(x_i^2-2x_iy_i+y_i^2)} \tag{3.3}$$

$$d(y,x) = \|y-x\|_2 = \sqrt{\sum_{i=1}^{n}(y_i-x_i)^2} = \sqrt{\sum_{i=1}^{n}(y_i^2-2y_ix_i+x_i^2)} \tag{3.4}$$

Nun erkennt man, dass die Gleichungen 3.3 und 3.4 die Gleichheit erfüllen.
zu (3): Mit Hilfe der Cauchy-Schwarz'schen Ungleichung lässt sich zeigen

$$d(x,y) = \|x-y\|_2 = \|x-z+z-y\|_2 \leq \|x-z\|_2 + \|z-y\|_2 = d(x,z)+d(z,y). \tag{3.5}$$

□

Somit kann man aufgrund der vorrangegangenen Definitionen eine Beschreibung der Cluster durchführen, welche dichte Punktwolken in einem n-dimensionalen Raum als Cluster zusammenfassen und durch Gebiete mit einer geringen Dichte getrennt werden. Werden aber die Bedingungen der Homogenität und Heterogenität nicht eingehalten, ist eine Clusteranalyse nicht sinnvoll. Dies wird durch die Abbildung 3.1 verdeutlicht. Aufgrund der Tatsache, dass in dieser Abbildung eine große Punktwolke dargestellt ist, macht es wenig Sinn eine Clusterunterteilung vorzunehmen.

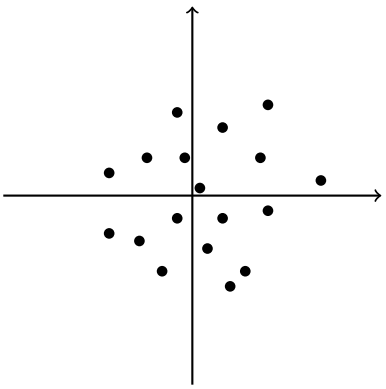


Abbildung 3.1: keine Struktur zu erkennen

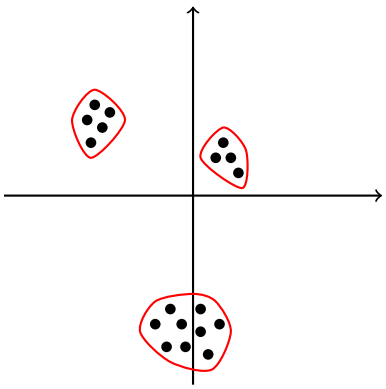


Abbildung 3.2: erkennbare Clusterstruktur

In der Abbildung 3.2 dagegen sind drei Cluster zu erkennen. Die Cluster erfüllen die Bedingung der Homogenität und die Bedingung der Heterogenität. Ein weiterer wichtiger Punkt ist die Skalierung der Wertebereiche. Dies wird erforderlich, um eine Vergleichbarkeit der Werte zu erreichen. In den Abbildungen 3.3 und 3.4 wird dies verdeutlicht, was auch in [HK13] genauer erläutert wird.

Die Abbildung 3.3 zeigt vier Punkte, die in die beiden Cluster $\{a,b\}$ und $\{c,d\}$ eingeteilt werden können. Die Abbildung 3.4 zeigt die gleichen Punkte mit einer anderen Skalierung der Achsen. Dabei entstehen die beiden Cluster $\{a,c\}$ und $\{b,d\}$. Dieser Effekt kann eine falsche Auswertung der Daten zur Folge haben. Darüber hinaus können

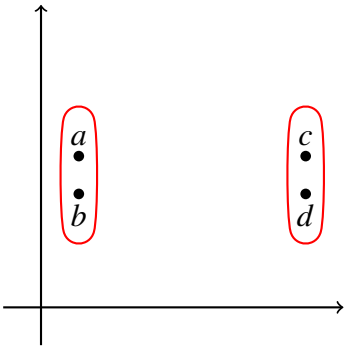


Abbildung 3.3: Punkte in Clusterunterteilung, vor Skalierung

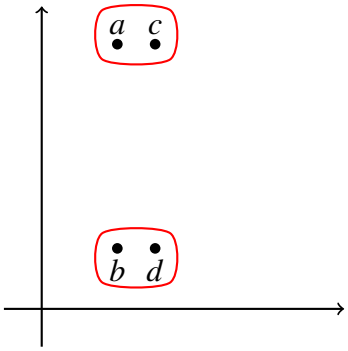


Abbildung 3.4: Punkte in Clusterunterteilung, nach Skalierung

auch nicht reellwertige Daten oder abstrakte Klassen, z.B. Währungen: Euro, Dollar, Yen usw., auftreten. Man kann abstrakten Klassen einen ganzzahligen Wert zuordnen. Aufgrund der Daten und der Fragestellung kann dies den gewünschten Effekt widerspiegeln, kann aber auch das Gegenteil eintreten lassen. Durch die Vergabe von Wertzuweisungen können falsche Informationen in die Daten hinein interpretiert werden. Zum Beispiel, dass die abstrakte Klasse mit der Nummer 1 eine größere Ähnlichkeit zur Klasse 2 besitzt als zur Klasse 4, obwohl eine solche Interpretation der Daten nicht sinnvoll ist.

Im weiteren Verlauf wird eine Einordnung der verschiedenen Verfahren der Clusteranalyse vorgenommen. Grundsätzlich kann man Clusteranalyseverfahren in unvollständige, deterministische und probabilistische Verfahren unterteilen, siehe [WKM11] und [BPW10].

Die unvollständigen Clusteranalyseverfahren werden in der Literatur auch geometrische Methoden genannt oder auch als Repräsentations- oder Projektionsverfahren bezeichnet. Solche Verfahren stützen sich darauf, die Daten nicht in Cluster zu unterteilen. Es wird nur eine Visualisierung der Daten in niedrigdimensionalen Räumen vorgenommen. Im Idealfall handelt es sich bei den niedrigdimensionalen Räumen um einen 1- oder 2-dimensionalen Raum, teilweise auch der 3-dimensionale Raum.

Verfahren der Clusteranalyse die als deterministisch bezeichnet werden, ordnen den jeweiligen Objekt eine Wahrscheinlichkeit von 0 oder 1 zu, ob diese zu Cluster gehören. Des Weiteren lassen sich deterministische Verfahren unterteilen in:

- Nichtüberlappende bzw. überlappungsfreie Verfahren. Dabei werden die Cluster so gebildet, dass jedes Objekt in den Daten genau einem Cluster angehört.
- Überlappende Verfahren. Die Cluster werden so gebildet, dass Objekte mehreren Clustern angehören können.

Die probabilistischen Verfahren kann man als Verallgemeinerung der deterministischen

Verfahren interpretieren. Die Objekte bekommen nicht nur mit einem Wert von 0 oder 1 zugewiesen, sondern mit einer Wahrscheinlichkeit zwischen 0 und 1. Demzufolge kann ein Objekt z.B. mit einer Wahrscheinlichkeit von 60% einem Cluster angehören.

Abbildung 3.5: Darstellung der verschiedenen Verfahren mit bekannten Algorithmen

In der Abbildung 3.5 werden die verschiedenen Verfahren graphisch dargestellt.

4 Algorithmus für die Ausreißererkennung

Ein Algorithmus für die Ausreißererkennung ist der DBSCAN, der in [EKSX96] vorgestellt wird. Die Abkürzung DBSCAN steht für Density Based Spatial Clustering of Applications with Noise, was übersetzt bedeutet Dichtebasierte räumliche Clusteranalyse mit Rauschen. Wie durch den Namen impliziert wird, wird die „Dichte“ in einer Punktmenge überprüft. Im Folgenden werden die einzelnen Merkmale bzw. Charakteristika beschrieben.

Im DBSCAN-Algorithmus wird überprüft, ob ein Punkt q innerhalb einer ε-Umgebung von p liegt, dazu muss der Abstand von p und q kleiner als ε sein. In der rechten Abbildung befinden sich zwei Punkte, die diese Bedingung erfüllen.

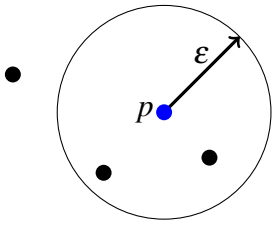


Abbildung 4.1: ε-Umgebung eines Punktes p

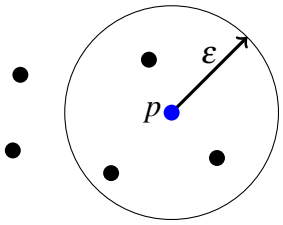


Abbildung 4.2: $minPts$

Der zweite Parameter $minPts$ legt fest, wie viele Punkte sich mindestens innerhalb der ε-Umgebung von dem Punkt p befinden müssen, damit dieser als Kernobjekt bezeichnet wird. Wenn diese Bedingung erfüllt ist, wird der Punkt p als Kernobjekt bezeichnet. Wenn in der linken Abbildung eine Anzahl der $minPts = 2$ festgelegt wird, ist p ein Kernobjekt.

Ein Punkt p heißt direkt-dichte-erreichbar von einem Punkt q, falls q ein Kernobjekt und p in der ε-Umgebung von q ist. In der rechten Abbildung ist q ein Kernobjekt, für $minPts = 2$. Der Punkt p befindet sich in der ε-Umgebung von q und ist somit direkt-dichte-erreichbar.

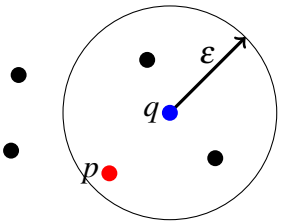


Abbildung 4.3: direkt-dichte-erreichbar

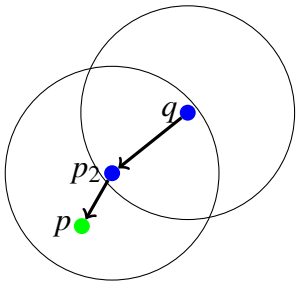


Abbildung 4.4: dichte-erreichbar

Ein Punkt p heißt dichte-erreichbar vom Punkt q, falls eine Kette von Punkten $q = p_1, \ldots, p_n = p$ existiert, sodass p_{i+1} direkt-dichte-erreichbar von p_i ist für alle $i = 1, ..., n-1$. In der linken Abbildung seien die Punkte q und p_2 Kernobjekte, somit bilden die Punkte $q = p_1, p_2, p_3 = p$ eine Kette und p ist dichte-erreichbar vom Punkt q.

Zwei Punkte t_1 und t_2 sind dichte-verbunden zu einem Punkt p, wenn t_1 und t_2 über p dichte-erreichbar sind. Man kann auch sagen, t_1 und t_2 sind über p dichte-verbunden. In der rechten Abbildung sind die Punkte t_1 und t_2 über den Punkt p dichte-verbunden.

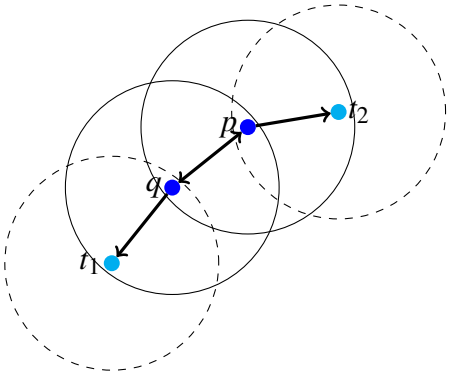


Abbildung 4.5: dichte-verbunden

Ein Cluster ist somit eine Menge von dichte-verbundenen Punkten, die maximal in Bezug auf die Dichte-Erreichbarkeit ist. Punkte, die zu keinem Cluster gehören, werden als Rauschpunkte bezeichnet. Somit sind Ausreißer bei der Messung nach John Rauschpunkte.

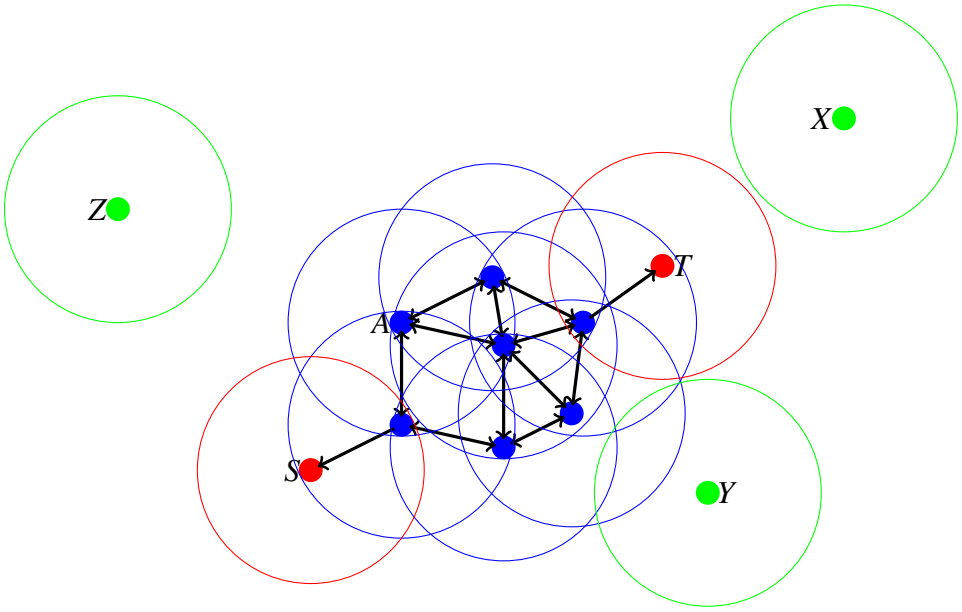


Abbildung 4.6: Darstellung des DBSCAN Algorithmus

In der Abbildung 4.6 ist die visuelle Darstellung des Algorithmus zusammengefasst. Die

blauen Punkte zählen zu den Kernobjekten, die roten Punkte sind dichte-erreichbar und die grünen Punkte gehören zu den Rauschpunkten. Die Punkte S und T sind über A dichte-verbunden und gehören deswegen zum gleichen Cluster. Die minimale Anzahl der Punkte in der ε-Umgebung ist in diesem Beispiel $minPts = 3$.

Der Algorithmus 1 und der Algorithmus 2 sind zusammen ein Pseudoquellcode für den DBSCAN Algorithmus. Der Pseudoquellcode stammt aus [Grä15].

Algorithm 1 DBSCAN-Algorithmus

```
1:  function DBSCAN(D, ε, minPts)
2:      C = 0
3:      for all nicht besuchten Punkte P in D do
4:          markiere P als besucht
5:          N = D.Gebietsabfrage(P, ε)
6:          if |N| < minPts then
7:              markiere P als Rauschpunkt
8:          else
9:              C = nächster Cluster
10:             expandCluster(P, N, C, ε, minPts)
```

Algorithm 2 expandCluster-Algorithmus

```
1:  function EXPANDCLUSTER(P, N, C, ε, minPts)
2:      füge P zum Cluster C hinzu
3:      for all Punkte P' in N do
4:          if P' nicht besucht then
5:              markiere P' als besucht
6:              N' = D.Gebietsabfrage(P', ε)
7:              if N' ≥ minPts then
8:                  N = N angeschlossen mit N'
9:          if P' noch nicht in irgendeinem Cluster then
10:             füge P' zum Cluster C dazu
```

5 Algorithmen für die Partitionierung

In diesem Kapitel wird auf die Partitionierung von Datenmengen eingegangen. Aufgrund der Anwendung für den Beugetest, wird hierbei die Anzahl der Cluster auf zwei bzw. drei festgelegt, siehe Kapitel 7 Anwendung der Clusteralgorithmen auf die Messdaten.

5.1 Expectation-Maximization-Algorithmus

Der Expectation-Maximization-Algorithmus, kurz EM-Algorithmus, gehört zu der latenten Profilanalyse, welche zu der Oberklasse der probabilistischen Verfahren zählt. Das Verfahren wird von Johann Bacher in [BPW10] und von Miin-Shen Yang in [YLL12] beschrieben. Für den Algorithmus werden folgende Modellannahmen getroffen:

1. Die Anzahl der Cluster ist K.
2. Die Cluster besitzen Anteilswerte von $\pi(k)$ an der Grundgesamtheit.
3. Jeder Cluster k besitzt in jedem Repräsenten des Clusters j eine Normalverteilung mit dem Clusterzentrum m_{kj} und der Varianz σ^2_{kj}.
4. In jedem Cluster sind die Datenpunkte i und j unabhängig.

Die Normalverteilung kann durch einen empirisch beobachten Wert x_{gj} im Cluster k beschrieben werden, der sich aus einem Fehlerterm ε_{gj} und dem Clusterzentrum μ_{kj} zusammensetzt.

$$x_{gj} = \mu_{kj} + \varepsilon_{gj}$$

Die zufälligen Messungenauigkeiten werden als normalverteilte Zufallsgröße mit Erwartungswert 0 und der Varianz σ^2_{kj} angenommen, deren Realisierung den Fehlerterm ε_{gj} erbigt. Aufgrund dessen, werden die Fehler verschiedener Messungen als unabhängig angenommen. Die im Folgenden aufgeführten Werte lassen sich auf Grund dieser Modellannahmen aufstellen:

$$\text{Mittelwert} \quad m_j = \sum_k \pi(k) \cdot \mu_{kj} \tag{5.1}$$

Der Mittelwert ist somit eine gewichtete Mittelung aller Mittelwerte.

$$\text{Kovarianz} \quad \sigma(i,j) = \sum_k \pi(k) \cdot (\mu_{ki} - \mu_i) \cdot (\mu_{kj} - \mu_j) \tag{5.2}$$

Die Kovarianz hängt nur von von den Abweichungen von den Clusterzentren ab.

$$\text{Varianz} \quad \sigma_j^2 = \sigma_{jj} = \sum_k \pi(k) \cdot \sigma_{kj}^2 + \sum_k \pi(k) \cdot (\mu_{kj} - \mu_j)^2 \tag{5.3}$$

Die Varianz hängt zusätzlich von Fehlerstreuungen ab. Um die Modellparameter bestmöglich zu schätzen, wird auf die Maximum-Likelihood-Methode zurückgegriffen.

Definition 5.1 (Maximum-Likelihood-Funktion) Die Maximum-Likelihood-Funktion für eine konkrete Stichprobe x aus der Verteilung X mit den unbekannten Parameter Θ ist gegeben durch

- $L(x,\Theta) = \prod_{i=1}^{n} p(x_i,\Theta)$, falls X diskret mit den Einzelwahrscheinlichkeiten $p(x_i,\Theta) = P(X = x_i,\Theta)$ oder
- $L(x,\Theta) = \prod_{i=1}^{n} f(x_i,\Theta)$, falls X stetig mit der Dichte $f(x_i,\Theta)$

definiert.

Somit ist im diskreten bzw. stetigen Fall $L(x,\Theta)$ die Wahrscheinlichkeit bzw. die Dichte dafür, dass genau die konkrete Stichprobe $x = (x_1,...,x_n)$ beobachtet wird, wenn Θ der unbekannte Parameter ist. Für die Schätzung der Parameter ist die Funktion

$$L = \prod_g \sum_k \pi(k) \cdot \pi(g|k) \tag{5.4}$$

oder der Logarithmus

$$l = ln(L) = \sum_g ln \sum_k \pi(k) \cdot \pi(g|k) \tag{5.5}$$

zu maximieren. Dabei ist $\pi(k)$ der Anteilswert des Clusters k und $\pi(g|k)$ die bedingte Wahrscheinlichkeit des Auftretens des Objektes (Vektors) g in dem Cluster k. Diese bedingte Wahrscheinlichkeit ist, aufgrund der Unabhängigkeit (siehe Modellannahme), gleich dem Produkt der bedingten Auftrittswahrscheinlichkeiten $\pi(x_{gi}|k)$ des Wertes von g in der Variablen j für den Cluster k.

$$\pi(g|k) = \prod_j \pi(x_{gj}|k) \tag{5.6}$$

Bei dem EM-Algorithmus ist die Auftrittswahrscheinlichkeit $\pi(x_{gj}|k)$ der Wert der Funktion der Normalverteilung mit dem Mittelwert μ_{kj} und der Varianz σ_{kj}^2.

$$\pi(x_{gj}|k) = \varphi(x_{gj}|\mu_{kj},\sigma_{kj}^2) = \frac{1}{\sigma_{kj} \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\frac{(x_{gj}-\mu_{kj})^2}{\sigma_{kj}^2}} \tag{5.7}$$

Die Zuordnungswahrscheinlichkeiten $\pi(k|g)$, mit der der Cluster k bei den Objekten g auftritt, werden als bekannt vorausgesetzt. Daraus folgt nun aus der Maximum-Likelihood-Funktion:

$$\begin{aligned} l &= \sum_g \sum_k \pi(k|g) \cdot ln\left[\pi(k) \cdot \pi(g|k)\right] \\ &= \sum_g \sum_k \pi(k|g) \cdot \left[ln(\pi(k)) + ln(\pi(g|k))\right] \\ &= \sum_g \sum_k \pi(k|g) \cdot ln(\pi(k)) + \pi(k|g) \cdot ln(\pi(g|k)) \\ &= \sum_g \sum_k \left[\pi(k|g) \cdot ln(\pi(k)) + \pi(k|g) \cdot ln\left(\prod_j \pi(x_{gj}|k)\right)\right] \\ &= \sum_g \sum_k \pi(k|g) \cdot ln(\pi(k)) + \sum_g \sum_k \sum_j \pi(k|g) \cdot ln(\pi(x_{gj}|k)) \end{aligned}$$

Somit teilt sich die Schätzaufgabe in die Schätzung von den Anteilswerten und die Schätzung der Parameter der Normalverteilung auf. Die Schätzwerte lassen sich wie folgt berechnen:

$$\pi(k) = P(k) = \frac{1}{n} \sum_g \pi(k|g) \tag{5.8}$$

$$\bar{x}_{kj} = \frac{\sum_g \pi(k|g) \cdot x_{gj}}{\sum_g \pi(k|g)} \tag{5.9}$$

$$s^2_{kj} = \frac{\sum_g \pi(k|g) \cdot \left(x_{gj} - \bar{x}_{kj}\right)^2}{\sum_g \pi(k|g)} \tag{5.10}$$

Die im EM-Algorithmus verwendeten Parameter k und g beschreiben jeweils die Anzahl der vorliegenden Cluster beziehungsweise die Punkte (Objekte). $P(k|g)$ ist die Wahrscheinlichkeit, dass ein Punkt g einem Cluster k angehört und wird als Zuordnungswahrscheinlichkeit bezeichnet.$P(g|k)$ ist die bedingte Wahrscheinlichkeit des Auftretens des Punktes g im Cluster k. Die Clusteranteilswerte werden durch $P(k)$ beschrieben. Jedes Cluster k besitzt in jeder Klassifikationsvariable j eine Normalverteilung mit dem Clusterzentrum bzw. Clustermittelwert $\bar{x}_{kj}$ und der Varianz s^2_{kj}.

Ablauf:

1. Eingabe oder Berechnung der Startwerte
2. Berechnung der Zuordnungswahrscheinlichkeiten:

$$P(k|g)^{(i)} = \frac{P(k)^{(i-1)} P(g|k)^{(i-1)}}{\sum_k P(k)^{(i-1)} P(g|k)^{(i-1)}}$$

wobei $P(g|k)^{(i-1)} = \prod_j P(x_{gj}|k)^{(i-1)} = \prod_j \phi\left(x_{gj}|\bar{x}_{kj}^{(i-1)}, s_{kj}^{2\ (i-1)}\right)$ ist und das hochgestellte i der Iterationszähler ist.

3. Neuberechnung der Modellparameter:

$$P(k)^{(i)} = \frac{1}{n}\sum_g P(k|g)^{(i)}$$

$$\bar{x}_{kj}^{(i)} = \frac{\sum_g P(k|g)^{(i)} x_{gj}}{\sum_g P(k|g)^{(i)}}$$

$$s_{kj}^{2\ (i)} = \frac{\sum_g P(k|g)^{(i)} \left(x_{gj} - \bar{x}_{kj}^{(i)}\right)^2}{\sum_g P(k|g)^{(i)}}$$

4. Prüfung der Konvergenz: Der Algorithmus bricht ab, wenn im dritten Schritt eine Verbesserung kleiner einem vorgegeben Wert, z.B. 10^{-6}, ist oder die maximale Abweichung der aufeinander folgenden Schätzwerte kleiner als einem bestimmten Wert, z.B. 10^{-3} ist.

Um zu zeigen, dass der EM-Algorithmus konvergiert, benötigt man Definition 5.2 und Satz 5.3. Der Beweis stammt aus der Vorlesung im Wintersemester 2014/15 aus dem Modul Computational Intelligence II der Fachhochschule Mittweida.

Definition 5.2 (konkave Funktion) Eine Funktion $f : I \to \mathbb{R}$ auf einem Intervall $I = [a,b] \subseteq \mathbb{R}$ heißt konkav, wenn

$$f(x+\lambda(y-x)) = f((1-\lambda)x+\lambda y) \geq (1-\lambda)f(x)+\lambda f(y)$$

für alle $\lambda \in [0,1]$ und alle $x, y \in I$ gilt.

Theorem 5.3 (Jensen Ungleichung) *Für jede konkave Funktion f, beliebige Zahlen x_i auf einem Intervall I und positive Zahlen λ_i mit $\sum_{i=1}^{n} \lambda_i = 1$, gilt*

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \geq \sum_{i=1}^{n} \lambda_i f(x_i).$$

Beweis:
Der Beweis für die Jensen Ungleichung erfolgt über vollständige Induktion. Für den Fall $n = 1$ entsteht die Gleichung, welche der Definition der Konkavität entspricht.

$$f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda)f(x_2)$$

Fü den Induktionsschluss von n auf $n+1$ werden folgende Notationen eingeführt:

$$\lambda := \lambda_1 + \lambda_2 + ... + \lambda_n = \sum_{i=1}^{n} \lambda_i \tag{5.11}$$

$$x := \frac{\lambda_1}{\lambda} x_1 + \frac{\lambda_2}{\lambda} x_2 + ... + \frac{\lambda_n}{\lambda} x_n = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} x_i \tag{5.12}$$

Daraus folgt

$$\lambda \cdot x = \sum_{i=1}^{n} \lambda_i x_i \text{ und } \lambda + \lambda_{n+1} = \lambda_1 + \lambda_2 + ... + \lambda_n + \lambda_{n+1} = 1 \tag{5.13}$$

und

$$\frac{\lambda_1}{\lambda} + \frac{\lambda_2}{\lambda} + ... + \frac{\lambda_n}{\lambda} = \frac{\lambda_1 + \lambda_2 + ... + \lambda_n}{\lambda} = \frac{\lambda}{\lambda} = 1 \tag{5.14}$$

Aus der Definition der Konkavität und der Induktionsannahme gilt:

$$\begin{aligned} f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &= f\left(\sum_{i=1}^{n} \lambda_i x_i + \lambda_{n+1} x_{n+1}\right) \\ &= f\left(\lambda \left(\sum_{i=1}^{n} \frac{\lambda_i}{\lambda} x_i\right) + \lambda_{n+1} x_{n+1}\right) \\ &= f(\lambda x + \lambda_{n+1} x_{n+1}) \\ &\geq \lambda f(x) + \lambda_{n+1} f(x_{n+1}) \\ &= \lambda f\left(\sum_{i=1}^{n} \frac{\lambda_i}{\lambda} x_i\right) + \lambda_{n+1} f(x_{n+1}) \end{aligned} \tag{5.15}$$

Wegen der Gleichung 5.11 und der Induktionsannahme folgt nun

$$f\left(\sum_{i=1}^{n} \frac{\lambda_i}{\lambda} x_i\right) \geq \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} f(x_i) \tag{5.16}$$

Somit folgt:

$$\begin{aligned} f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &\geq \lambda f\left(\sum_{i=1}^{n} \frac{\lambda_i}{\lambda} x_i\right) + \lambda_{n+1} f(x_{n+1}) \\ &\geq \lambda \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} f(x_i) + \lambda_{n+1} f(x_{n+1}) \\ &= \sum_{i=1}^{n+1} \lambda_i f(x_i) \end{aligned} \tag{5.17}$$

□

Die Jensen Ungleichung liefert nun die Folgerung, da der Logarithmus konkav ist und für beliebige positive Zahlen c_k gilt:

$$log\left(\sum_k c_k\right) = log\left(\sum_k \lambda_k \frac{c_k}{\lambda_k}\right) \geq \sum_k \lambda_k log\left(\frac{c_k}{\lambda_k}\right) \tag{5.18}$$

Des Weiteren will man die Kostenfunktion oder auch Energiefunktion maximieren.

$$K(X,\Theta) = \sum_{i=1}^{n} log\left(\sum_{j=1}^{m} g(x_i,\Theta_j)\right) \longrightarrow max \tag{5.19}$$

Die Funktion $g(x_i,\Theta_j)$ ist eine positiv, beschränkte Funktion. Laut der Folgerung entsteht die Funktion

$$\sum_i log\left(\sum_j g(x_i,\Theta_j)\right) \geq \sum_i \sum_j \lambda_{ij} log\left(\frac{g(x_i,\Theta_j)}{\lambda_{ij}}\right) =: \mathscr{L}(\Lambda,\Theta,X)$$

$\mathscr{L}(\Lambda,\Theta,X)$ ist eine untere Schranke für die Kostenfunktion, die für positive λ_{ij} mit $\sum_j \lambda_{ij} = 1$ für alle $i = 1,...,n$. Demzufolge lässt sich die Kostenfunktion wie folgt aufschreiben.

$$K(X,\Theta) = \mathscr{L}(\Lambda,\Theta,X) + \mathscr{K}(\Lambda,\Theta,X)$$

Somit ist eine Darstellung der Funktion $\mathscr{K}$ gesucht. Dafür sei $p_{ij} := \frac{g(x_i,\Theta_j)}{\sum_k g(x_i,\Theta_k)}$, sodass $\sum_j p_{ij} = 1$ für alle i und $p_{ij} \geq 0$ für alle i und j, Wahrscheinlichkeiten sind. Dann gilt für die Gleichung

$$K(X,\Theta) = \mathscr{L}(\Lambda,\Theta,X) + \mathscr{K}(\Lambda,\Theta,X) \tag{5.20}$$

mit

$$\mathscr{L}(\Lambda,\Theta,X) = \sum_i \sum_j \lambda_{ij} log\left(\frac{g(x_i,\Theta_j)}{\lambda_{ij}}\right) \tag{5.21}$$

$$\mathscr{K}(\Lambda,\Theta,X) = \sum_i \sum_j \lambda_{ij} log\left(\frac{\lambda_{ij}}{p_{ij}}\right) \tag{5.22}$$

Aufgrund der Tatsache, dass die Funktion $\mathscr{K}$ den Bedingungen der Kullback-Leibler-Divergenz genügt, siehe [RAC04],

$$\mathscr{K}(\Lambda,\Theta,X) \geq 0$$

$$\mathscr{K}(\Lambda,\Theta,X) = 0 \Leftrightarrow \lambda_{ij} = p_{ij} \text{ für alle } i,j$$

reicht es die Richtigkeit der angegebenen Funktion $\mathscr{K}$ zu zeigen. Dafür muss man diese nur in die Gleichung 5.20 einsetzen.

$$\begin{aligned}
\mathscr{L}(\Lambda,\Theta,X)+\mathscr{K}(\Lambda,\Theta,X) &= \sum_i\sum_j \lambda_{ij} log\left(\frac{g(x_i,\Theta_j)}{\lambda_{ij}}\right)+\sum_i\sum_j \lambda_{ij} log\left(\frac{\lambda_{ij}}{p_{ij}}\right)\\
&= \sum_i\left[\sum_j \lambda_{ij} log\left(\frac{g(x_i,\Theta_j)}{\lambda_{ij}}\right)+\sum_j \lambda_{ij} log\left(\frac{\lambda_{ij}}{p_{ij}}\right)\right]\\
&= \sum_i\left[\sum_j \lambda_{ij} log\left(\frac{g(x_i,\Theta_j)}{\lambda_{ij}}\right)+\sum_j \lambda_{ij} log\left(\frac{\lambda_{ij}\sum_k g(x_i,\Theta_k)}{g(x_i,\Theta_j)}\right)\right]\\
&= \sum_i\left[\underbrace{\sum_j \lambda_{ij}}_{=1} log\left(\sum_k g(x_i,\Theta_k)\right)\right]\\
&= \sum_i log\left(\sum_k g(x_i,\Theta_k)\right)\\
&= K(X,\Theta)
\end{aligned}$$

Die Vorgehensweise des EM-Algorithmus sieht vor, dass am Anfang zufällige Startvektoren (K_{start}) ausgewählt werden. Im nächsten Schritt werden die Wahrscheinlichkeiten p_{ij} berechnet und $\lambda_{ij} = p_{ij}$ gesetzt. Durch diesen Schritt erhält die Funktion $\mathscr{K}(\Lambda,\Theta,X)$ den Wert 0, was zur Folge hat, dass nun die Gleichung $K(X,\Lambda) = \mathscr{L}(\Lambda,\Theta,X)$ entsteht. Im letzten Schritt wird Λ festgehalten und neue Vektoren Θ bestimmt, sodass $\mathscr{L}(\Lambda,\Theta^{neu},X) \geq \mathscr{L}(\Lambda,\Theta^{alt},X)$ gilt. Dies wurde in der rechten Abbildung 5.1 grafisch dargestellt.

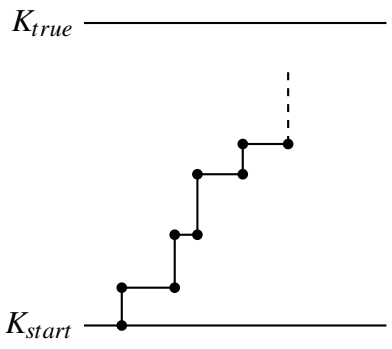


Abbildung 5.1: Konvergenz

Aufgrund dieser Vorgehensweise nimmt der Algorithmus immer einen „besseren“ Wert an und verschlechtert sich nicht. In der Regel erreicht man nur ein lokales Optimum, zur Zeit ist kein solcher Algorithmus, der ein globales Optimum liefert, bekannt.

5.2 k-Means Algorithmus

Der k-Means Algorithmus dient zur Partitionierung der Daten in k Cluster. Dabei entsteht folgendes Optimierungsproblem, wobei die Summe der quadratischen Abweichungen

von den Clusterschwerpunkten minimal werden soll.

$$J = \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|_2 \longrightarrow min \tag{5.23}$$

Dabei sind die x_j die Datenpunkte und μ_i die Schwerpunkte der Cluster S_i. Die Grundidee dieser Minimierung basiert auf der Methode der kleinsten Quadrate und wird aus diesem Grund auch Clustering durch Varianzminimierung genannt.

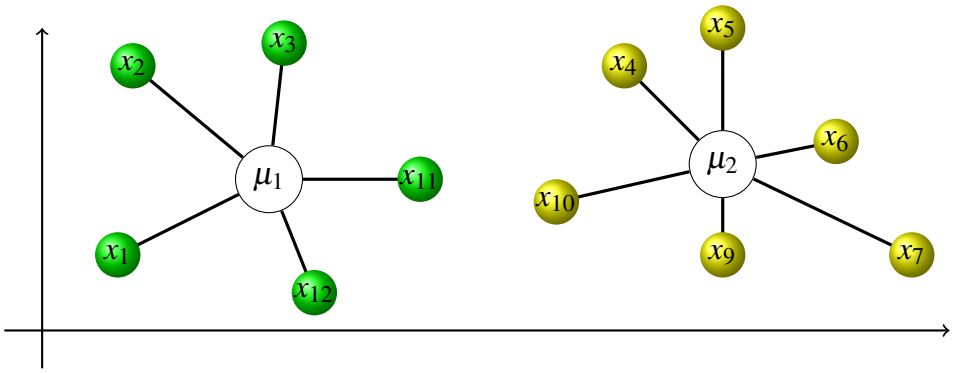


Abbildung 5.2: visuelle Darstellung der Abstände zum Clustermittelpunkt

Da dieses Problem zur Klasse der NP-schweren Problemen gehört, gibt es keinen Algorithmus der dieses Problem in polynomialer Laufzeit löst, außer es gilt $P = NP$. Deswegen gibt es verschiedene Heuristiken, welche dieses Problem bewältigen. Der bekannteste Ansatz, um die Minimierungsaufgabe zu lösen, bietet der Lloyd Algorithmus. In der Literatur wird der Algorithmus von Lloyd als der Standard Algorithmus von k-Means verwendet, siehe [Cep15].

Initialisierung Wahl von k zufälligen Mittelwerten (Means bzw. Clusterzentren) $m_1^{(0)}, ..., m_k^{(0)}$ aus dem Datensatz

Zuordnung Jeder Datenpunkt wird dem Cluster zugeordnet, bei dem der Abstand zum Clustermittelpunkt am kleinsten ist

$$S_i^{(t)} = \{x_j : \|x_j - m_i^{(t)}\|_2 \leq \|x_j - m_{i^*}^{(t)}\|_2 \text{ für alle } i^* = 1, ..., k\}$$

Aktualisierung Neuberechnung der Mittelpunkte der Cluster

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

Es wird die Zuordnung und Aktualisierung solange durchgeführt, bis sich die Zuordnung nicht mehr ändert oder zyklisch ist.

In den folgenden Abbildungen ist der Durchlauf für den k-Means Algorithmus mit $k = 3$ Clustern, exemplarisch dargestellt. Im ersten Schritt (siehe Abb. 5.3) sind die Datenpunkte als graue Quadrate gekennzeichnet. Da dieser Datensatz in drei Cluster aufgeteilt werden soll, werden zufällig drei Clusterzentren bestimmt, welche als blauer, roter und oranger Kreis dargestellt sind. Im nächsten Schritt (siehe Abb. 5.4) werden die grauen Datenpunkte den jeweiligen nächsten Clusterzentrum zugeordnet. Im Schritt 3 (siehe Abb. 5.5) werden die Clusterzentren neu berechnet und an die neue Position verschoben. Der letzte Schritt (siehe Abb. 5.6) sieht vor, dass die Datenpunkte wieder den jeweiligen Clusterzentren zugeordnet werden, dessen Zentrum am nächsten ist. Der Vorgang von Schritt 2 bis Schritt 4 wird solange wiederholt, bis keine Neuanordnung der Cluster stattfindet.

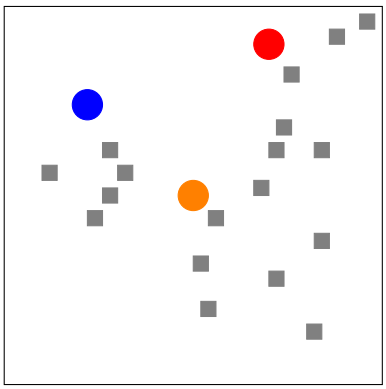


Abbildung 5.3: Schritt 1

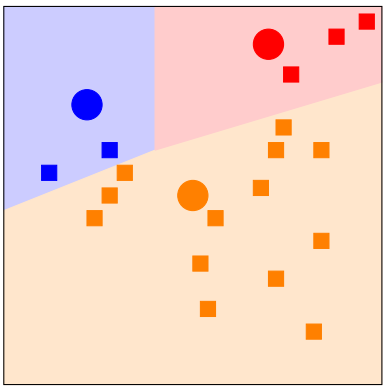


Abbildung 5.4: Schritt 2

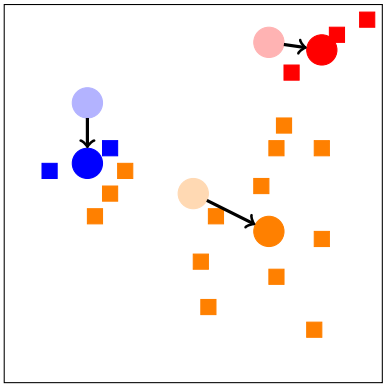


Abbildung 5.5: Schritt 3

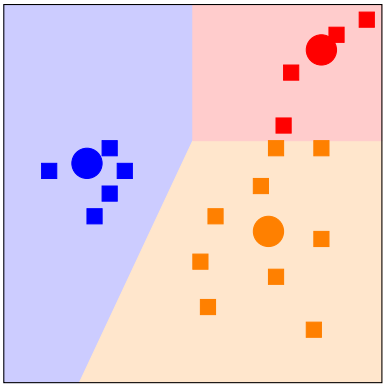


Abbildung 5.6: Schritt 4

Im Folgenden wird die Konvergenz des k-Means Algorithmus gezeigt. Dazu werden weitere Notationen benötigt.

$S_1^{(t)}, S_2^{(t)}, ..., S_k^{(t)}$ - die einzelnen Cluster in der t-ten Iteration

$m_1^{(t)}, m_2^{(t)}, ..., m_k^{(t)}$ - die einzelnen Mittelwerte in der t-ten Iteration

Da die Partitionierung einer endlichen Menge eine endliche Anzahl von nöglichen Partitionen zur Folge hat reicht es zu zeigen, dass die Kostenfunktion monoton fallend ist.

$$J(S_1,...,S_k,m_1,...,m_k) = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - m_i\|_2 \tag{5.24}$$

Da die Funktion monoton fallend sein soll, muss gelten:

$$J\left(S_1^{(t+1)},...,S_k^{(t+1)},m_1^{(t+1)},...,m_k^{(t+1)}\right) \leq J\left(S_1^{(t)},...,S_k^{(t)},m_1^{(t)},...,m_k^{(t)}\right) \tag{5.25}$$

Aufgrund der Tatsache, dass durch die Aufteilung der Vektoren auf die nächsten Mittelwerte keine Mehrkosten entstehen, gilt die Ungleichung:

$$J\left(S_1^{(t+1)},...,S_k^{(t+1)},m_1^{(t)},...,m_k^{(t)}\right) \leq J\left(S_1^{(t)},...,S_k^{(t)},m_1^{(t)},...,m_k^{(t)}\right) \tag{5.26}$$

Nun muss die Richtigkeit der nächsten Ungleichung gezeigt werden.

$$J\left(S_1^{(t+1)},...,S_k^{(t+1)},m_1^{(t+1)},...,m_k^{(t+1)}\right) \leq J\left(S_1^{(t+1)},...,S_k^{(t+1)},m_1^{(t)},...,m_k^{(t)}\right) \tag{5.27}$$

Um dies zu beweisen, muss gezeigt werden das durch das arithmetische Mittel die Kostenfunktion eines Clusters S mit dem Mittelwert m

$$J(S,m) := \sum_{x \in S} \|x - m\|^2 = \sum_{x \in S} \sum_{j=1}^{n} (x_j - m_j)^2 \tag{5.28}$$

minimiert wird.

$$\begin{aligned} \frac{\partial}{\partial m_i} J(S,m) &\stackrel{!}{=} 0 \\ -2 \sum_{x \in S} (x_i - m_i) &= 0 \\ \Leftrightarrow \sum_{x \in S} x_i - |C| m_i &= 0 \\ \Leftrightarrow m_i &= \frac{1}{|C|} \sum_{x \in S} x_i \end{aligned}$$

Somit wurde gezeigt, dass die Kostenfunktion durch das arithmetische Mittel minimiert wird und die Ungleichung korrekt ist. Demzufolge löst der Algorithmus das Minimierungsproblem, in der Regel wird aber nur ein lokales Optimum gefunden.

5.3 Fuzzy c-Means Algorithmus

Im Gegensatz zum k-Means Algorithmus wird beim Fuzzy c-Means den Datenpunkten eine Wahrscheinlichkeit zugeordnet mit welcher der Datenpunkt zum Cluster gehört. Frank Höppner beschreibt dies in [HK13] und wird auch im Artikel von James Bezdek

[BEF84] näher erläutert. Die Gleichung 5.29 beschreibt die Zielfunktion des Fuzzy c-Means Algorithmus.

$$F = \sum_{i=1}^{c} \sum_{j=1}^{k} u_{ij}^{m} \|v_i - x_j\|_2^2 \longrightarrow min \tag{5.29}$$

In der Zielfunktion wurde eine sogenannte Fuzzifier Ziffer $m > 1$ eingeführt. Damit wird verhindert, dass die Zugehörigkeitsgrade u_{ij} nur die Werte 0 oder 1 annehmen. $(v_1, ..., v_c)$ ist die Matrix der Clusterprototypen. Die Zielfunktion 5.29 ist unter den folgenden Nebenbedingungen 5.30 und 5.31 zu minimieren.

$$\sum_{i=1}^{c} u_{ij} = 1 \quad \forall j = 1, ..., k \tag{5.30}$$

$$u_{ij} \geq 0 \quad \forall i \in \{1, ..., c\},\ j \in \{1, ..., k\} \tag{5.31}$$

Aufgrund der Nebenbedingung 5.30 spricht man von einer probabilistischen Clusteranalyse. Man interpretiert die Zugehörigkeitsgrade als Wahrscheinlichkeiten. Die folgenden beiden Bedingungen ergeben sich als notwendige Bedingungen für das Minimum.

$$v_i = \frac{\sum_{j=1}^{k} u_{ik}^{m} \cdot x_j}{\sum_{j=1}^{k} u_{ik}^{m}} \tag{5.32}$$

$$u_{ij} = \frac{1}{\sum_{l=1}^{c} \left(\frac{\|v_i - x_j\|_2^2}{\|v_l - x_j\|_2^2} \right)^{\frac{1}{m-1}}} \tag{5.33}$$

Beweis der Gleichung 5.32:
Da alle Richtungsableitungen von der Funktion 5.29 nach v_i notwendig 0 sein müssen, gilt für alle $\xi \in \mathbb{R}^n$ mit $t \in \mathbb{R}$:

$$\begin{aligned}
0 &= \frac{\partial}{\partial v_i} \sum_{j=1}^{k} \sum_{i=1}^{c} u_{ji}^{m} \|x_j - v_i\|_2^2 \\
&= \sum_{j=1}^{k} u_{ij}^{m} \frac{\partial}{\partial v_i} \|x_j - v_i\|_2^2 \\
&= \sum_{j=1}^{k} u_{ij}^{m} \lim_{t \to 0} \frac{\|x_j - (v_i + t\xi)\|_2^2 - \|x_j - v_i\|_2^2}{t} \\
&= \sum_{j=1}^{k} u_{ij}^{m} \lim_{t \to 0} \frac{\big(((x_j - v_i) - t\xi)^T ((x_j - v_i) - t\xi) - (x_j - v_i)^T (x_j - v_i)\big)}{t} \\
&= \sum_{j=1}^{k} u_{ij}^{m} \lim_{t \to 0} \frac{-2t(x_j - v_i)^T \xi + t^2 \xi^T \xi}{t}
\end{aligned}$$

$$= -2 \sum_{j=1}^{k} u_{ij}^m \lim_{t \to 0} (x_j - v_i)^T \xi$$

$$\Leftrightarrow \sum_{j=1}^{k} u_{ij}^m (x_j - v_i) = 0$$

$$\Leftrightarrow v_i = \frac{\sum_{j=1}^{k} u_{ik}^m \cdot x_j}{\sum_{j=1}^{k} u_{ik}^m}$$

□

Der Beweis der Gleichung 5.35:
Mit der Kostenfunktion 5.29 und der Nebenbedingung

$$\sum_{i=1}^{c} u_{ij} = 1 \quad \text{für alle } j$$

stellt man die Lagrange Funktion

$$L = \sum_{i=1}^{c} \sum_{j=1}^{k} u_{ij}^m \|v_i - x_j\|_2^2 - \sum_{j=1}^{k} \lambda_j \left(\left(\sum_{i=1}^{c} u_{ij} \right) - 1 \right) \tag{5.34}$$

auf, von welcher nun die partiellen Ableitungen gleich Null gesetzt werden müssen.

$$\frac{\partial L}{\partial u_{ij}} = m \cdot u_{ij}^{m-1} \cdot \|v_i - x_j\|_2^2 - \lambda_j \stackrel{!}{=} 0 \tag{5.35}$$

$$\frac{\partial L}{\partial \lambda_j} = \left(\sum_{i=1}^{c} u_{ij} \right) - 1 \stackrel{!}{=} 0 \tag{5.36}$$

Die Gleichung 5.35 lässt sich nun nach u_{ij} umstellen.

$$u_{ij} = \left(\frac{\lambda_j}{m \cdot \|v_i - x_j\|_2^2} \right)^{\frac{1}{m-1}} \tag{5.37}$$

Die nach u_{ij} umgestellte Gleichung kann somit in die Gleichung 5.36 eingesetzt werden.

$$\sum_{l=1}^{c} \left(\frac{\lambda_j}{m \cdot \|v_l - x_j\|_2^2} \right)^{\frac{1}{m-1}} = 1$$

$$\left(\frac{\lambda_j}{m} \right)^{\frac{1}{m-1}} \sum_{l=1}^{c} \left(\frac{1}{\|v_l - x_j\|_2^2} \right)^{\frac{1}{m-1}} = 1$$

$$\left(\frac{\lambda_j}{m}\right)^{\frac{1}{m-1}} = \frac{1}{\sum_{l=1}^{c}\left(\frac{1}{\|v_l - x_j\|_2^2}\right)^{\frac{1}{m-1}}} \tag{5.38}$$

Aufgrund einer Umstellung der Gleichung 5.35 nach $\frac{\lambda_j}{m}$, lässt sich dies in die Gleichung 5.38 einsetzten.

$$\begin{aligned}
\left(u_{ij} \cdot \|v_i - x_j\|_2^2\right)^{\frac{1}{m-1}} &= \frac{1}{\sum_{l=1}^{c}\left(\frac{1}{\|v_l - x_j\|_2^2}\right)^{\frac{1}{m-1}}} \\
u_{ij} \cdot \|v_i - x_j\|_2^{2\frac{1}{m-1}} &= \frac{1}{\sum_{l=1}^{c}\left(\frac{1}{\|v_l - x_j\|_2^2}\right)^{\frac{1}{m-1}}} \\
u_{ij} &= \frac{1}{\sum_{l=1}^{c}\left(\frac{1}{\|v_l - x_j\|_2^2}\right)^{\frac{1}{m-1}}} \cdot \frac{1}{\|v_i - x_j\|_2^{2\frac{1}{m-1}}} \\
u_{ij} &= \frac{1}{\sum_{l=1}^{c}\left(\frac{\|v_i - x_j\|_2^2}{\|v_l - x_j\|_2^2}\right)^{\frac{1}{m-1}}}
\end{aligned}$$

□

Der Algorithmus setzt sich nun wie folgt zusammen.

Ablauf

1. Erstelle die Matrix $U^{(0)}$, bestehend aus den Wahrscheinlichkeiten u_{ij} für alle i, j.
2. Berechnung der Prototypen $V^{(t)} = [v_i]$ im t-ten Iterationsschritt
3. Berechnung der Matrix $U^{(t)} = [u_{ij}]$ im t-ten Iterationsschritt
4. Falls $\|U^{(t)} - U^{(t-1)}\| < \varepsilon$ dann Stopp, sonst mit Schritt 2 fortfahren.

6 Bewertung von Clusterverfahren und deren Clusterbildung

Nach allem was in den beiden vorherigen Kapiteln beschrieben ist, stellt sich die Frage, ob die Verfahren korrekt für dieses Problem sind, welches in der Vorlesung von [JKS+13] präsentiert wird. Dabei ist zu beachten, dass ein Cluster, welcher mittels eines der vorgestellten Clusterverfahren erzeugt wurde, nicht als richtig oder falsch zu bewerten ist. Man spricht davon, ob der Cluster sinnvoll oder weniger sinnvoll ist. Dabei basieren sinnvolle Clusterverfahren auf einer qualitativ guten Annahme bzw. Heuristik. Um Aussagen über die Sinnhaftigkeit zu treffen, muss eine gewisse Kenntnis über die Datenmenge voraus gesetzt werden. In der Abbildung 6.1 bis 6.6 ist dies verdeutlicht.

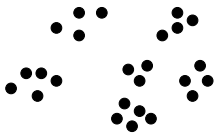


Abbildung 6.1: Originaldaten

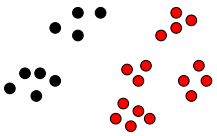


Abbildung 6.2: zwei Cluster

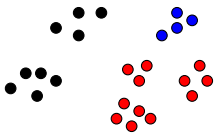


Abbildung 6.3: drei Cluster

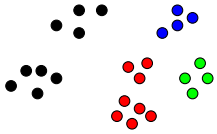


Abbildung 6.4: vier Cluster

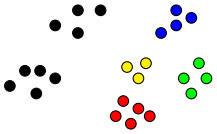


Abbildung 6.5: fünf Cluster

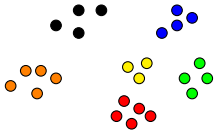


Abbildung 6.6: sechs Cluster

Es ist dabei zu beachten, dass es verschiedene Möglichkeiten gibt eine Menge von Punkten zu clustern (siehe Definiton 3.2).

Definition 6.1 (Kompaktheit eines Clusters nach [Reh14]) Ein Maß für die Kompaktheit eines Clusters C mit dem Clusterzentrum m ist definiert durch

$$TD(C) = \sum_{x \in C} \|x - m\|_2.$$

Definition 6.2 (Kompaktheit eines Clusterings nach [Reh14]) Ein Maß für die Kompaktheit eines gesamten Clusterings mit k Clustern $C_1, ..., C_k$ ist definiert durch

$$TD = \sum_{i=1}^{k} TD(C_i).$$

Definition 6.3 (Silhouette nach [Reh14]) Die Silhouette ist ein Gütemaß für die Bewertung von Partitionierungsverfahren, bei dem die Anzahl der Cluster unabhängig vom

Maß ist. Sie wird definiert durch:

$$s(p) = \frac{b(p) - a(p)}{max\{a(p), b(p)\}}.$$

Dabei ist $a(p)$ der Abstand eines Punktes p zum Clusterzentrum des Punktes p und $b(p)$ ist der Abstand eines Punktes p zum nächstgelegenden fremden Clusterzentrum.

Die Silhouette liegt im Intervall $[-1, 1]$. Für diesen Wert gelten folgende Eigenschaften:

$s(p) < 0$	Der Punkt p liegt näher an einem fremden Cluster, gehört aber zu einem anderen Cluster. Daraus folgt, dass das Clustering verbessert werden kann.
$s(p) \approx 0$	Der Punkt p liegt zwischen seinem Cluster und dem nächstgelegendem Cluster.
$s(p) > 0$	Der Punkt p befindet sich im selbigen Cluster.

Definition 6.4 (Silhouettenkoeffizient nach [Reh14]) Der Silhouettenkoeffizient für ein Cluster C ist definiert durch

$$s_C = \frac{1}{|C|} \sum_{p \in C} s(p).$$

7 Anwendung der Clusteralgorithmen auf die Messdaten

In diesem Kapitel werden die einzelnen Verfahren der Clusteranalyse auf die Messdaten angewandt.

Für den DBSCAN-Algorithmus wurden folgende zwei Tabellen aufgeührt, um zu verdeutlichen, was passiert, wenn man einen der beiden Parameter $(\varepsilon, minPts)$ festhält und den anderen frei wählt. Um einen Berechnungsvorschlag der Parameter $(\varepsilon, minPts)$ aus den Daten bereitzustellen, werden folgende Größen benötigt. Die Anzahl in der Messung erhobenen Daten wird mit n festgelegt und p ist ein bestimmter Prozentsatz, der vom Anwender festgelegt wird.

$$minPts = \lfloor n \cdot p \rfloor$$

Für die ε-Umgebung wird der durchschnittliche Weg sd von Punkt zu Punkt berechnet.

$$sd = \frac{1}{n-1} \sum_{i=1}^{n-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$$

Daraus folgt eine Berechnung für die ε-Umgebung mit

$$\varepsilon = sd \cdot a.$$

Der Faktor a wird in die Berechnung miteinbezogen, weil somit eine zu kleine ε-Umgebung verhindert wird. In der Tabelle 7.1 werden die $minPts$ mit $p = 0,5\%$ festgelegt und die ε-Umgebung verändert. Es ist in der Tabelle 7.1 zu erkennen, dass je größer die ε-

a	Epsilon	$minPts$	Anteil der Clusterpunkte
$0,9$	$0,193$	55	$0,02\%$
1	$0,214$	55	$0,02\%$
$1,1$	$0,236$	55	$84,97\%$
$1,2$	$0,257$	55	$87,49\%$
$1,3$	$0,279$	55	$91,19\%$
$1,4$	$0,300$	55	$92,89\%$

Tabelle 7.1: Anteil der geclusterten Datenpunkte mittels DBSCAN bei festen $minPts$ in Abhängigkeit von ε

Umgebung wird, desto größer der Anteil der Clusterpunkte wird. In der Tabelle 7.2 wird die ε-Umgebung festgehalten mit $a = 1,3$ und der Prozentsatz p verändert. Die Ergebnisse in der Tabelle 7.2 lassen darauf schließen, dass je kleiner der Prozentsatz p wird,

p	$minPts$	Epsilon	Anteil der Clusterpunkte
$0,1\%$	11	$0,279$	$93,49\%$
$0,2\%$	22	$0,279$	$91,19\%$
$0,3\%$	33	$0,279$	$91,19\%$
$0,4\%$	44	$0,279$	$91,19\%$
$0,5\%$	55	$0,279$	$91,19\%$
$0,6\%$	66	$0,279$	$91,19\%$
$0,7\%$	77	$0,279$	$91,19\%$
$0,8\%$	88	$0,279$	$0,02\%$

Tabelle 7.2: Anteil der geclusterten Datenpunkt mittels DBSCAN bei festen ε in Abhängigkeit von $minPts$

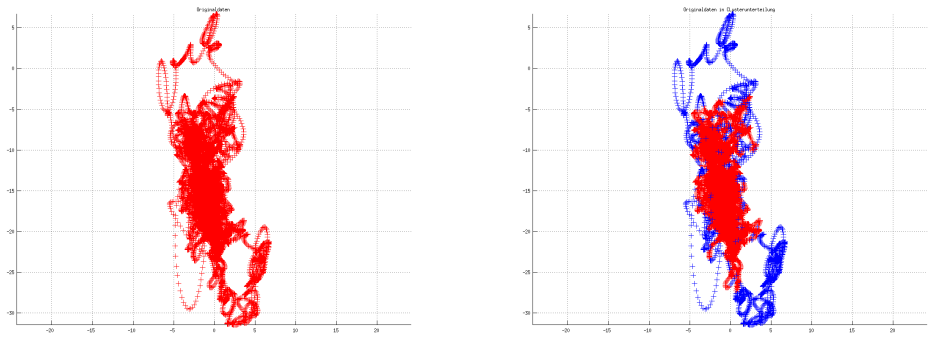
desto größer der Anteil der Clusterpunkte wird. Im Allgemeinen würde es sich für den Anwender anbieten die ε-Umgebung zu verändern, da dadurch eine genauere Kalibrierung erfolgen kann.

Um geeignete Startparameter festzulegen, wurden in einer Messreihe folgende Resultate erzeugt. Die Messreihe umfasst eine Gruppe von 8 Probanden die jeweils über einen Zeitraum von fünf Tagen jeweils eine Messung am Tag durchgeführt haben.

Durchgeführte Messung	a	p
ohne BalancePad, Augen geöffnet	$1,3$	$0,4\%$
ohne BalancePad, Augen geschlossen	$1,4$	$0,3\%$
mit BalancePad, Augen geöffnet	$1,5$	$0,2\%$
mit BalancePad, Augen geschlossen	$1,6$	$0,1\%$

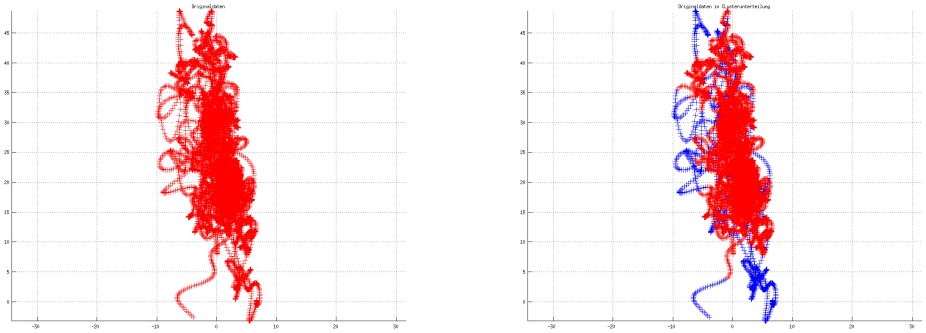
Tabelle 7.3: Startwerte für die Messungen nach John

Diese Werte lieferten gute Ergebnisse und sind in der Regel eine gute Startlösung. Hierbei entstanden folgende Bilder:



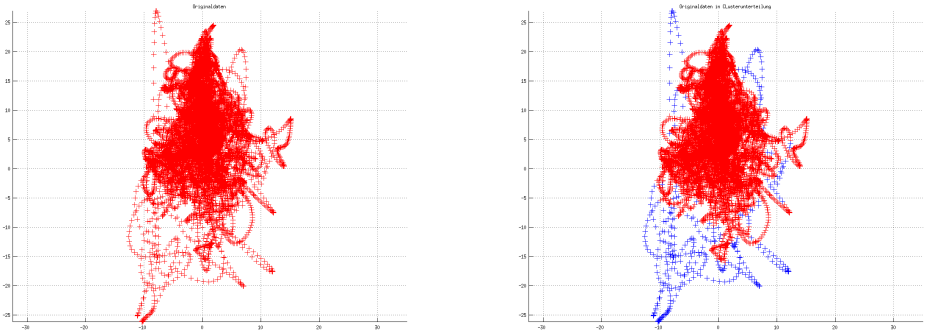
(a) Originaldaten (b) Originaldaten nach dem Clustern

Abbildung 7.1: 1. Messung nach John



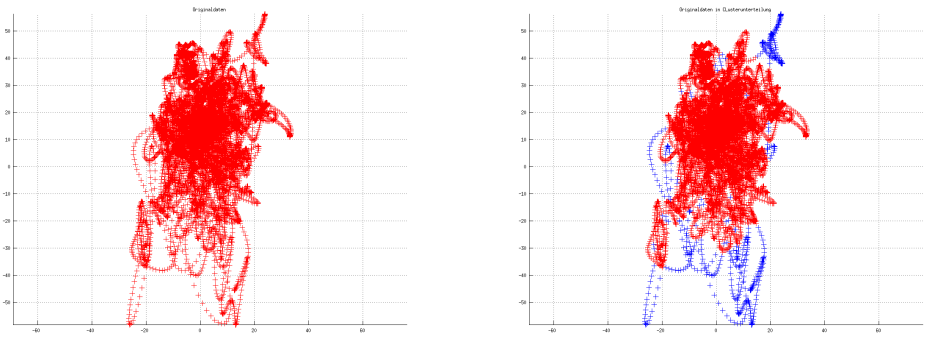
(a) Originaldaten (b) Originaldaten nach dem Clustern

Abbildung 7.2: 2. Messung nach John



(a) Originaldaten (b) Originaldaten nach dem Clustern

Abbildung 7.3: 3. Messung nach John



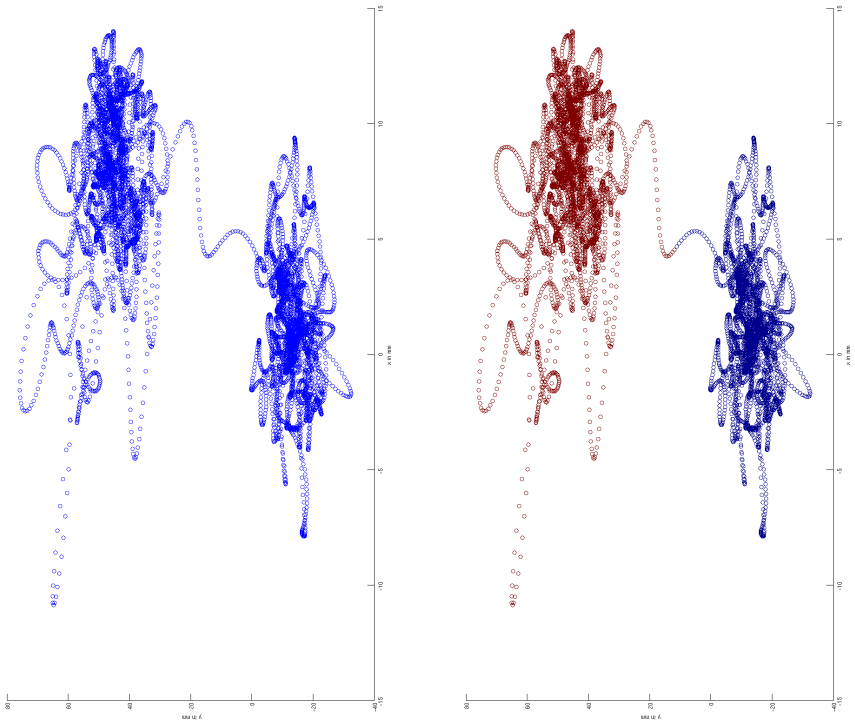
(a) Originaldaten (b) Originaldaten nach dem Clustern

Abbildung 7.4: 4. Messung nach John

Eine Anwendung der Algorithmen für die Partitionierung bietet sich im Beugetest. Dabei wird nur eine Bewegung aus der Grundhaltung nach vorne vollführt und diese gebeugte Haltung bis zum Schluss eingenommen. Dabei wird eine Clusteranzahl von zwei festgelegt und in der Abbildung 7.5 dargestellt.

Des Weiteren können auch auf den erweiterten Beugetest die gleichen Algorithmen angewandt werden, bei dem der Proband aus der Grundhaltung in den gebeugten Zustand übergeht und danach sich wieder in die Grundhaltung zurückbegibt. Aus diesem Grund erhalten die Daten eine dritte Dimension, die Zeit. Um die Clusteralgorithmen anwenden zu können, muss der Datensatz um eine Dimension reduziert werden. Die x-Ebene kann vernachlässigt werden, weil nur der zeitliche Abstand zwischen den Datenpunkten von Bedeutung ist und der Verlauf in der y-Ebene. Der Verlauf in der y-Ebene beschreibt den Bewegungsablauf des Nach-vorne-Beugens. Somit entstehen folgende Bilder 7.6.

Nun ist man in der Lage die dritte Dimension $(x$-Ebene$)$ wieder in den Datensatz aufzunehmen und somit das folgende Bilder 7.7 entstehen zu lassen für alle Algorithmen.



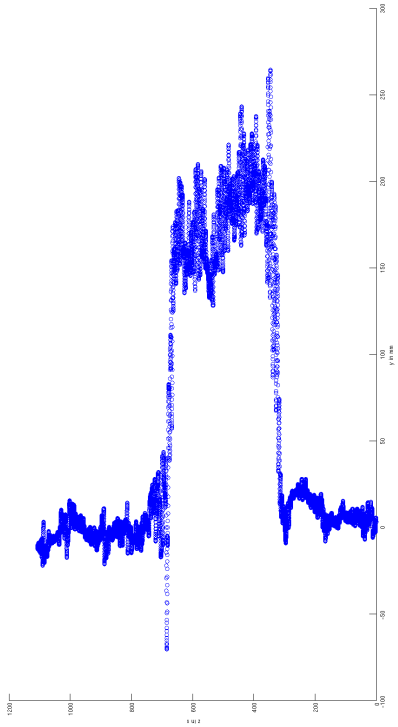
(a) Originaldaten

(b) EM Algorithmus angewandt

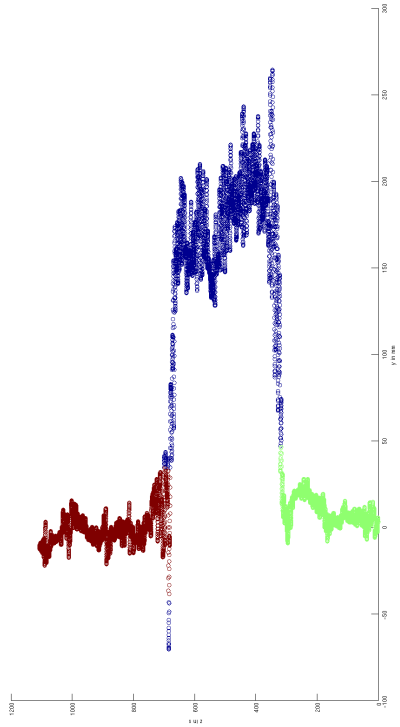
(c) k-Means Algorithmus angewandt

(d) Fuzzy c-Means Algorithmus angewandt

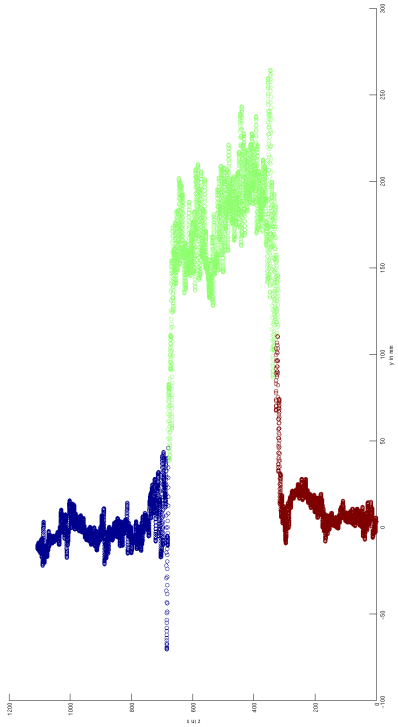
Abbildung 7.5: Beugetest



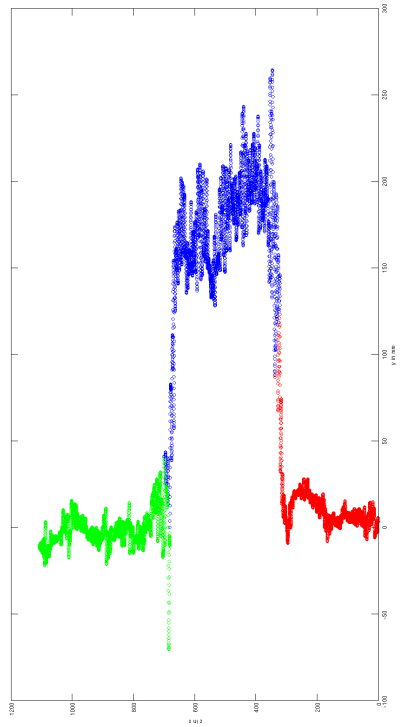
(a) Originaldaten



(b) EM Algorithmus angewandt

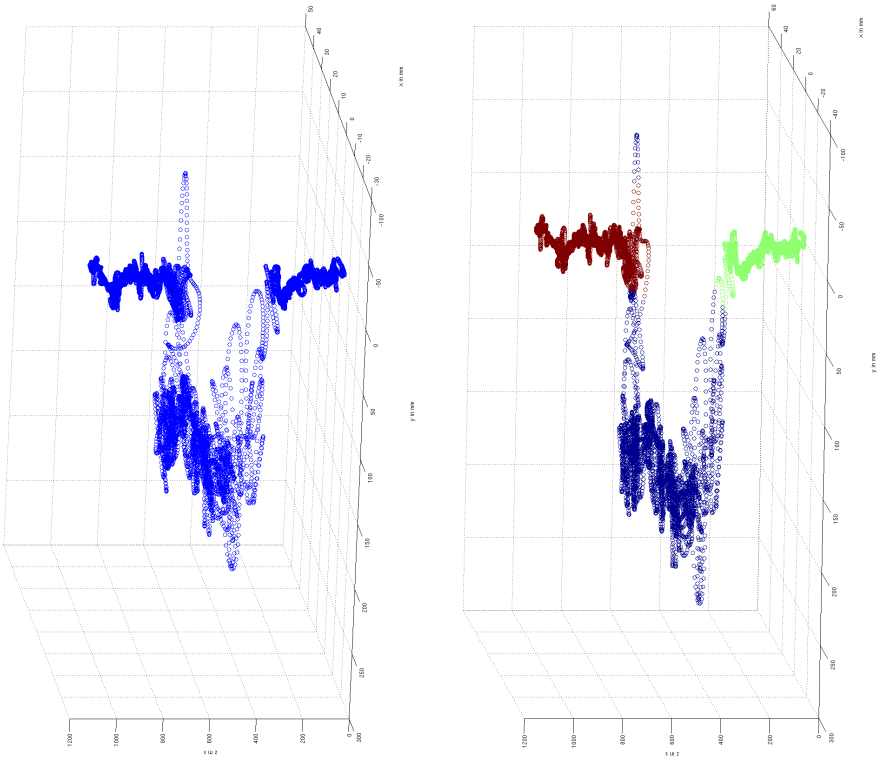


(c) k-Means Algorithmus angewandt



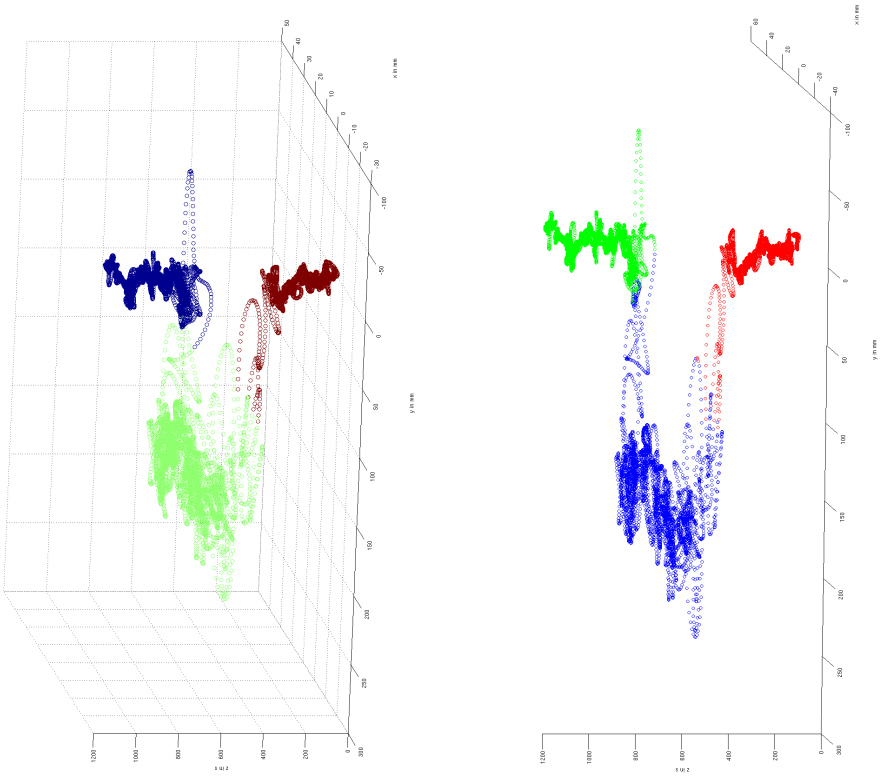
(d) Fuzzy c-Means Algorithmus angewandt

Abbildung 7.6: erweiterter Beugetest auf 2 Dimension reduziert



(a) Originaldaten

(b) EM Algorithmus angewandt



(c) k-Means Algorithmus angewandt

(d) Fuzzy c-Means Algorithmus angewandt

Abbildung 7.7: erweiterter Beugetest mit 3 Dimensionen

8 Statistische Kenngrößen

Die folgenden, aufgeführten statistischen Kenngrößen sind sowohl für die Auswertung des Beugetest und für die Messung nach John verwendbar. Die Anzahl der Messwerte der Originaldaten des Beugetest wird mit n abgekürzt und die Dauer der Messung mit t. Die Anzahl der Messpunkte im Cluster 1 des Beugetests wird mit n_1 bezeichnet. Für diesen Cluster wird die Dauer der Messung berechnet, mit der Gleichung $t_1 = \frac{t \cdot n_1}{n}$ in Sekunden. Für den Cluster 2 des Beugetests gilt analog n_2 als Anzahl der Messpunkte und $t_2 = \frac{t \cdot n_2}{n}$ als die Messdauer in Sekunden.
Für die Messung nach John gelten die gleichen Aussagen in Bezug auf die Anzahl der Messpunkte n bzw. der Dauer der Messung t.
Die im folgenden aufgezählten Formeln sind schreib-technisch für die Originaldaten der ersten Messung nach John aufgelistet, werden aber analog für den Beugetest beziehungsweise für die Cluster verwendet. Die Entfernung der Clustermittelpunkte, die Standardabweichung in x-Richtung, die Geradlinigkeit und der Quotient der Kreise sind statistische Kenngrößen, die speziell für den Beugetest konzipiert sind.

Eine Bewertung der Kenngrößen hinsichtlich des Gütegrades, d.h. ob der Proband mit seinen Werten im grünen Breich liegt, wird in dieser Bachelorarbeit nicht näher erläutert. Dies wurde im Praktikumsbericht von Christian Schulz durchgeführt. Dieser Bericht ist im Unternehmen Medizin & Service und bei Herrn Lindner an der Fachhochschule Mittweida hinterlegt und einsehbar.

8.1 Durchschnittlicher Messpunkt

Für den durchschnittlichen Messpunkt oder Mittelpunkt der Messwerte, wird für die x-Werte bzw. für die y-Werte jeweils das arithmetische Mittel gebildet.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Der Mittelpunkt hat die Koordinaten $M(\overline{X};\overline{Y})$.

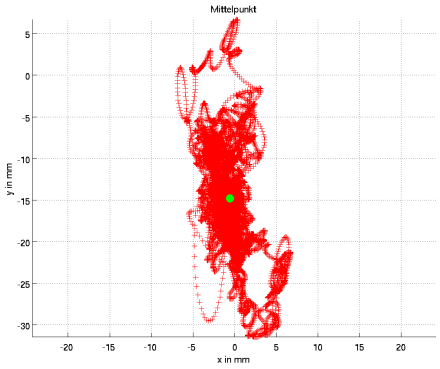


Abbildung 8.1: Mittelpunkt der Messwerte

$$M(-0,592\,mm; -14,828\,mm)$$

8.2 Varianz und Standardabweichung

Die Varianz ist die mittlere quadratische Abweichung vom Mittelpunkt der Messwerte. Durch das Quadrieren werden größere Abweichungen in der Messung stärker berücksichtigt. Dies kann dazuführen, dass Ausreißer stärker in die Berechnung eingehen.

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \qquad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$

Durch das Radizieren erhält man die Standardabweichung, welche nun die gleiche Einheit wie die Messwerte besitzt und somit besser für die Praxis interpretierbar ist.

$$s_X = \sqrt{s_X^2} \qquad s_Y = \sqrt{s_Y^2}$$

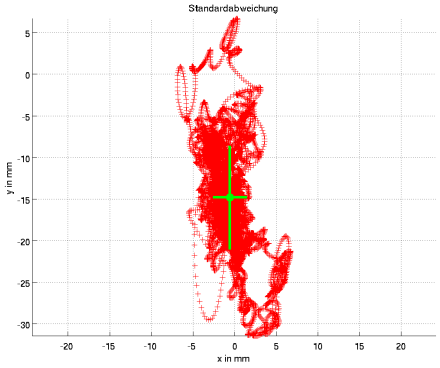


Abbildung 8.2: Standardabweichung

$$s_X = 2,031\,mm \qquad s_Y = 6,222\,mm$$

8.3 Variationskoeffizient

Der Variationskoeffizient bewertet das Größenverhältnis von der Standardabweichung und dem arithmetischen Mittel.

$$v_X = \frac{s_X}{\overline{X}} \qquad v_Y = \frac{s_Y}{\overline{Y}}$$

Durch diese Normierung der Varianz lassen sich Stichproben mit unterschiedlichem Mittelwert oder Standardabweichung vergleichen.

Für das Beispiel sind die Variationskoeffizienten $v_X = -3,432$ und $v_Y = -0,420$. Dieses Beispiel demonstriert, dass durch die Standardisierung der Varianzen, die x-Werte eine größere Streuung besitzen im Bezug zum arithmetischen Mittel als die y-Werte. Nur durch den Variationskoeffizient wird ein Vergleich zwischen den x und y Werten möglich.

8.4 AD-Streuung

Bei der AD-Streuung werden die Abstände aller Messpunkte vom Mittelpunkt betrachtet. Im Bezug auf die Standardabweichung ist die AD Streuung ein kleinerer Streuungswert, $s_X \geq AD_X$. AD ist eine Abkürzung und steht für „average deviation“, was übersetzt mittlere bzw. durchschnittliche Abweichung bedeutet. Die Einheit wird für die gemessenen Daten in mm angegeben.

$$AD_X = \frac{1}{n}\sum_{i=1}^{n}\left(\left|X_i - \overline{X}\right|\right) \qquad AD_Y = \frac{1}{n}\sum_{i=1}^{n}\left(\left|Y_i - \overline{Y}\right|\right)$$

Für den Test ist die AD Streuung ein besseres Maß, weil jeder Messpunkt mit der gleichen Gewichtung einbezogen wird. Im Gegenzug zur Varianz, werden bei der AD-Streuung durch den Betrag größere Werte nicht stärker berücksichtigt.

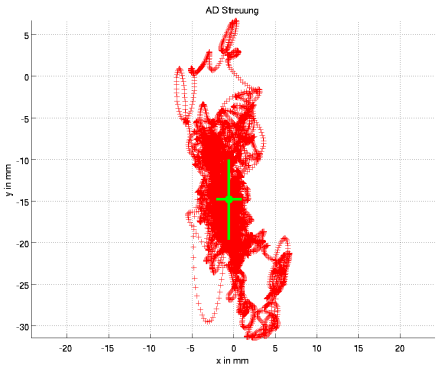


Abbildung 8.3: AD-Streuung

$$AD_X = 1,507\,mm$$

$$AD_Y = 4,797\,mm$$

8.5 Spannweite

Um die Spannweite der Messung zu ermitteln, benötigt man den kleinsten (X_{min} bzw. Y_{min}) Wert und den größten (X_{max} bzw. Y_{max}) Wert. Die Spannweite ist die Differenz vom größten und kleinsten Wert.

$$X_{span} = X_{max} - X_{min} \qquad Y_{span} = Y_{max} - Y_{min}$$

Die Spannweite dient zur Feststellung in welchem Bereich die Messung durchgeführt wurde.

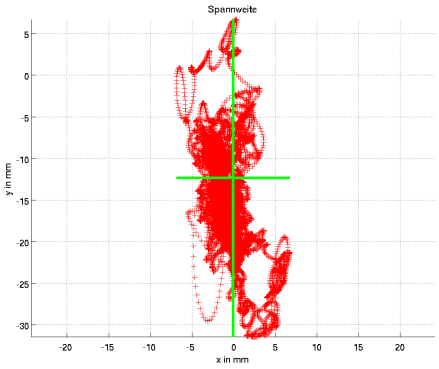


Abbildung 8.4: Spannweite

$$X_{span} = 13,61\,mm \quad Y_{span} = 38,14\,mm$$

8.6 Durchschnittsradius

Der Durchschnittsradius wird als Durchschnittswert der Abweichungen vom Mittelpunkt angesehen und wird bei den Daten in der Einheit mm angegeben. Die Berechnung erfolgt durch eine Bestimmung der Abstände vom Mittelpunkt, welche aufsummiert und durch die Gesamtanzahl dividiert werden. Somit entsteht ein Durchschnittswert, der die Abweichung der Messpunkte in jeder Richtung vom Mittelpunkt berücksichtigt.

$$R = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left(X_i - \overline{X}\right)^2 + \left(Y_i - \overline{Y}\right)^2}$$

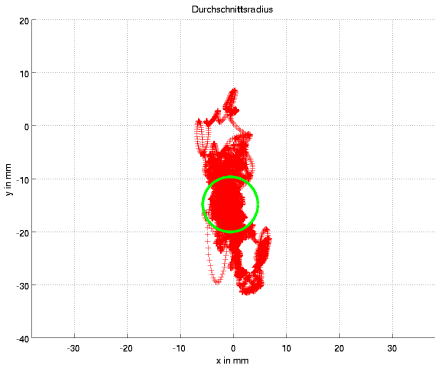


Abbildung 8.5: Durchschnittsradius

$$R = 5,21\,mm$$

8.7 Pendelweg

Der Pendelweg ist ein Maß für die Gesamtdistanz, den die Person auf der Messplatte zurückgelegt hat. Dabei wird der Abstand von Messpunkt zu Messpunkt berechnet und die einzelnen Abstände aufaddiert. Für die Daten ist die Einheit in m, weil die Zahlenwerte sonst zu groß sind.

$$S = \sum_{i=1}^{n-1} \sqrt{(X_{i+1} - X_i)^2 + (Y_{i+1} - Y_i)^2}$$

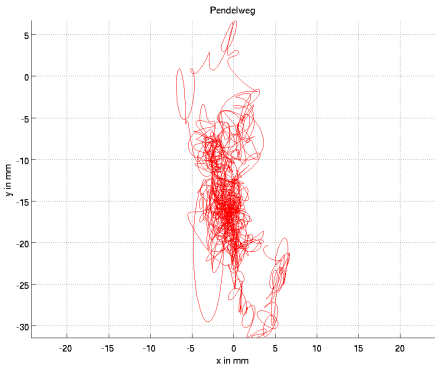


Abbildung 8.6: Pendelweg

$$S = 1,596 m$$

8.8 Durchschnittsgeschwindigkeit

Die Durchschnittsgeschwindigkeit ist der Quotient aus dem zurückgelegten Weg und der benötigten Zeit und wird in $\frac{mm}{s}$ oder $\frac{km}{h}$ angegeben. Dafür wird die Gesamtlänge des zurückgelegten Weges ermittelt (siehe Pendelweg) und die Gesamtzeit t.

$$D = \frac{S}{t} = \frac{1}{t} \sum_{i=1}^{n-1} \sqrt{(X_{i+1} - X_i)^2 + (Y_{i+1} - Y_i)^2}$$

Die Durchschnittsgeschwindigkeit für das Beispiel beträgt $D = 53,2 \frac{mm}{s} = 0,192 \frac{km}{h}$

8.9 Bewertungsmaß der Momentangeschwindigkeit

Um Aussagen über einen ruhigen Stand zu treffen oder ob der Stand instabil ist, wird eine Untersuchung der Momentangeschwindigkeiten erforderlich. Dabei werden die Geschwindigkeiten in der x-Richtung und in der y-Richtung berechnet. In der unteren Darstellung ist der Weg in x-Richtung mit s_1 und der Weg in y-Richtung mit s_2 bezeichnet. Der Weg in der jeweiligen Richtung wird durch den Zeitabstand zwischen den einzelnen Messpunkten geteilt. Dieser Zeitabstand berechnet sich durch $tt = \frac{t}{n}$. Nun erhält man die Geschwindigkeiten v_1 und v_2 in der x-Richtung beziehungsweise in der y-Richtung.

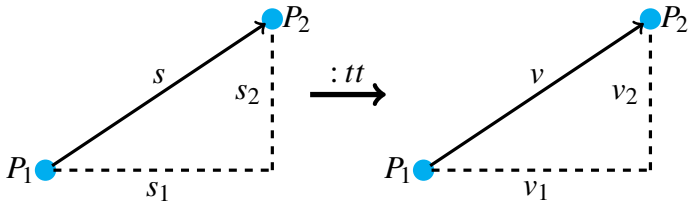
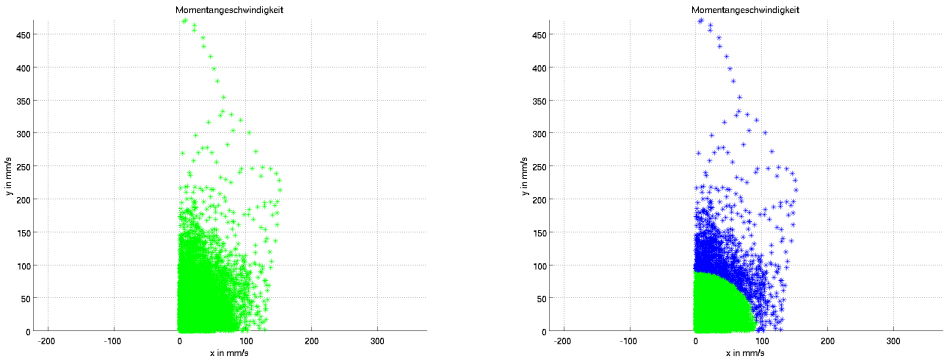


Abbildung 8.7: Darstellung der Berechnung der Momentangeschwindigkeit

Durch eine Festlegung einer Konstante c kann man den Anteil der Punkte bestimmen, für die gilt $v < c$. Diese Punkte kennzeichnen einen ruhigen Stand. Momentangeschwindigkeiten, für die $v \geq c$ gilt, charakterisieren einen unruhigen, nicht sicheren Stand. Die Festlegung der Konstante c erfolgt über die Bewertungstabelle der Durchschnittsgeschwindigkeit (siehe Abschnitt Durchschnittsgeschwindigkeit). Der Parameter, der den günstigen bzw. grünen Bereich charakterisiert, dient als Grundlage für die Konstante c, jeweils für die erste bis vierte Messung nach John.
In der Abbildung (a) sind die Originaldaten der Momentangeschwindigkeit dargestellt. Die (b) Abbildung zeigt die lokalen Geschwindigkeiten, die einen ruhigen Stand (grün) darstellen. Die Punkte, die außerhalb des grünen Bereiches (blau) liegen, sind Geschwindigkeiten größer als die Konstante c, die in diesem Zusammenhang für einen unruhigen Stand zu interpretieren sind.



(a) Originaldaten (b) Unterteilung in ruhig und unruhig

Abbildung 8.8: Maß der Momentangeschwindigkeit

Mittels dieser Festlegung der Konstante c auf die Momentangeschwindigkeiten, lassen sich die zu hohen Geschwindigkeiten, die auf einen unruhigen Stand hindeuten, im Bild der Originaldaten darstellen. Die blauen Sterne symbolisieren die Streckenabschnitte, wo die Person unstet oder unruhig stand.

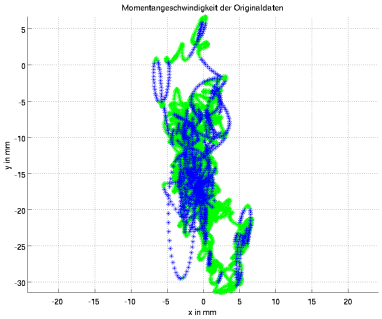


Abbildung 8.9: Darstellung der Momentangeschwindigkeit in den Daten

$$88,61\%$$

Das Bewertungsmaß der Momentangeschwindigkeit bezieht sich auf dem Anteil der Punkte, die grün gefärbt sind (also auf einen ruhigen Stand hindeuten) und wird in Prozent angegeben.

8.10 Durchschnittsschwankung

Die Durchschnittsschwankung ist eine mittlere Abweichung der Messwerte vom Mittelpunkt und wird in der Einheit mm angegeben. Sie wird berechnet als Wurzel der Varianz von der Summe der x-Koordinaten und der y-Koordinaten,
$K = \sqrt{var(X+Y)}$. Die Varianz als Summe von X und Y lässt sich wie folgt berechnen:

$$var(X+Y) = var(X) + var(Y) + 2 \cdot cov(X,Y)$$

Die Varianz von X ist gleich s_X^2 und die Varianz von Y ist gleich s_Y^2 (siehe Varianz bzw. Standardabweichung). Die Kovarianz von X und Y lässt sich durch

$$cov(X,Y) = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right) \left(Y_i - \overline{Y}\right)$$

berechnen.

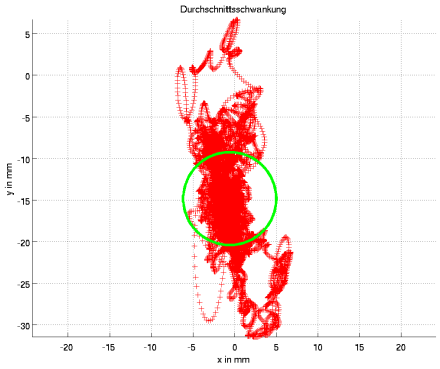


Abbildung 8.10: Durchschnittsschwankung

$$K = 5,589\,mm$$

Zusammengefasst entsteht die Gleichung der Durchschnittsschwankung:

$$K = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(\left(X_i - \overline{X}\right) + \left(Y_i - \overline{Y}\right)\right)^2}$$

8.11 Verweildauer der Quadranten

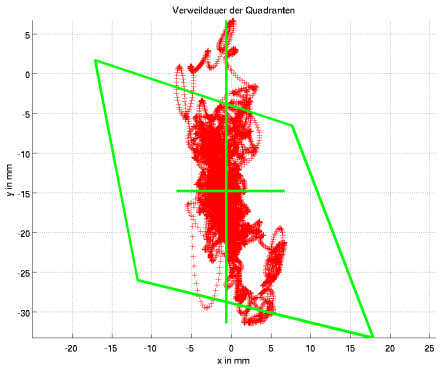


Abbildung 8.11: Verweildauer der Quadranten

Mit der Verweildauer der Quadranten wird der prozentuale Anteil der Messpunkte im $I.$, $II.$, $III.$ oder $IV.$ Quadranten im Koordinatensystem angegeben. Dabei wird eine Verschiebung der Koordinatenachsen zum Mittelpunkt der Messwerte vorgenommen. In der Abbildung ist dies durch die grünen Linien zu erkennen. Je gleichmäßiger der Anteil auf die einzelnen Quadranten verteilt ist, desto besser steht die Person auf der Messplatte.

$$I = 15,19\% \quad II = 30,38\%$$

$$III = 20,55\% \quad IV = 33,88\%$$

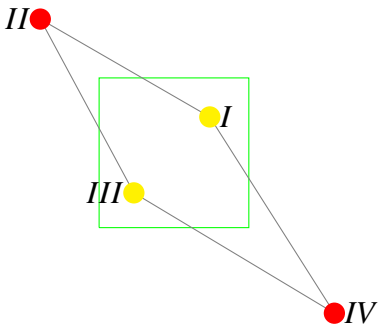


Abbildung 8.12: Visualisierung der Verweildauer der Quadranten

Eine weitere Möglichkeit der Darstellung der Verweildauer kann mittels zweier Vierecke vorgenommen werden. Das grüne Viereck ist ein Quadrat und soll den Idealfall symbolisieren, dass man in jedem Quadranten eine Verweildauer von 25% hat. Das zweite Viereck veranschaulicht die in der Messung berechneten Ergebnisse. Wenn die Punkte innerhalb des grünen Quadrats liegen bedeutet dies, dass man sich in diesem Quadranten zu wenig aufgehalten hat. Im Beispiel, wäre dies der $I.$ und $III.$ Quadrant. Wenn die Punkte sich außerhalb des Quadranten befinden, hat man sich zu lange in diesem Quadranten aufgehalten. In der rechten Abbildung sind es der $II.$ und $IV.$ Quadrant. Die Farben geben den Grad der Abweichung an. Ziel für den Probanden ist es, dass die Punkte in den jeweiligen Ecken verschoben werden.

8.12 Entfernung der Clustermittelpunkte

Die Entfernung der Clustermittelpunkte ist ein Maß für den Beugetest. Von den vorliegenden Clustern wird jeweils der Mittelpunkt $M_{Ruhe}(\overline{X}_{Ruhe};\overline{Y}_{Ruhe})$ des Ruhezustandes ermittelt und der Mittelpunkt $M_{Beuge}(\overline{X}_{Beuge};\overline{Y}_{Beuge})$ des Beugezustandes. Der Abstand berechnet sich wie folgt:

$$E = \sqrt{\left(\overline{X}_{Ruhe} - \overline{X}_{Beuge}\right)^2 + \left(\overline{Y}_{Ruhe} - \overline{Y}_{Beuge}\right)^2}$$

Die Einheit ist in mm.

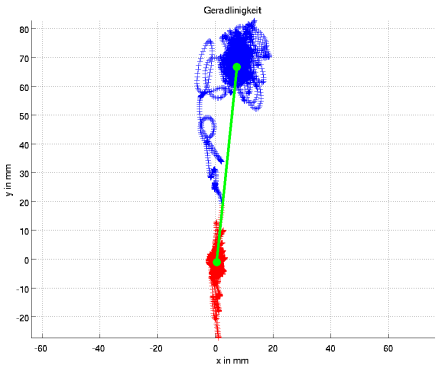


Abbildung 8.13: Entfernung der Clustermittelpunkte

$$E = 68,15\,mm$$

8.13 Standardabweichung in x-Richtung

Die Standardabweichung in x-Richtung ist ein Maß für die Bewertung des Beugetests. Es wird jeweils für den Cluster des Ruhezustandes und für den Cluster des Beugezustandes die Standardabweichung von x berechnet (siehe Varianz bzw. Standardabweichung). Die Einheit ist in mm.

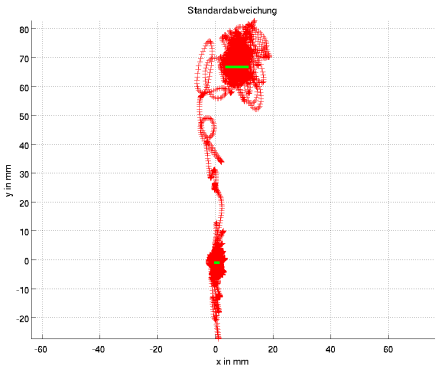


Abbildung 8.14: Standardabweichung in x-Richtung

$$s_{xRuhe} = 0,894\,mm \quad s_{xBeuge} = 3,294\,mm$$

8.14 Geradlinigkeit

Ein spezielles Maß für die Bewertung des Beugetests ist die Geradlinigkeit. Sie bewertet, wie gerade die Person sich nach vorne gebeugt hat und dabei diese Position ruhig einhalten kann. Sie entsteht aus der Berechnung einer linearen Funktion, unter Verwendung der beiden Clustermittelpunkte $M_{Ruhe}(\overline{X}_{Ruhe}, \overline{Y}_{Ruhe})$ und $M_{Beuge}(\overline{X}_{Beuge}, \overline{Y}_{Beuge})$. Der Anstieg der Funktion berechnet sich durch

$$m = \frac{\overline{Y}_{Ruhe} - \overline{Y}_{Beuge}}{\overline{X}_{Ruhe} - \overline{X}_{Beuge}}$$

und der Schnittpunkt mit der y-Achse berechnet sich durch

$$n = \overline{Y}_{Ruhe} - m * \overline{X}_{Ruhe}.$$

Somit erhält man die Geradengleichung $y = mx + n$.
Die Geradlinigkeit wird in Abhängigkeit des Anstieges berechnet:

$$G = 1 - \frac{1}{1 + |m|}$$

Je größer das Maß der Geradlinigkeit ist, desto besser kann sich die Person nach vorne beugen und diese Position ruhig halten. Dieses Maß wird in Prozent angegeben.

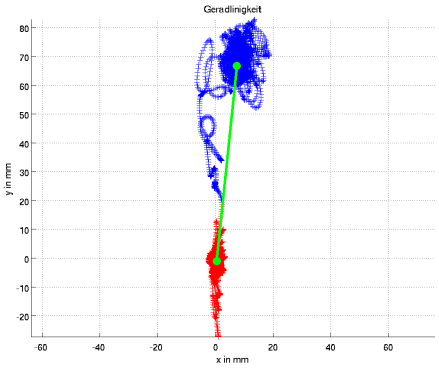


Abbildung 8.15: Geradlinigkeit

$$G = 96,3\%$$

8.15 Quotient der Kreise

Der Quotient der Kreise ist ein Maß für den Beugetest. Es werden jeweils die Durchschnittsradien des Ruheclusters und des Beugecluster berechnet (siehe Durchschnittsradius) und ins Verhältnis gesetzt.

$$Q = \frac{R_{Beuge}}{R_{Ruhe}}$$

Dieses Maß ist dimensionslos.

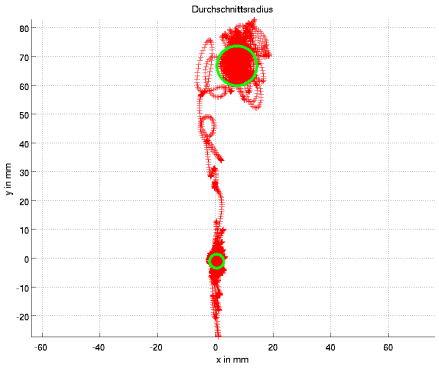


Abbildung 8.16: Quotient der Kreise

$$Q = 7,872$$

9 Zusammenfassung und Ausblick

Die verschiedene Algorithmen für die Bildung von Clusterings sollen einen Einblick in die Vielfalt der Clusteranalyse geben. Der DBSCAN-Algorithmus für die Ausreißererkennung wurde wegen seiner guten Heuristik ausgewählt und aufgrund seiner guten Ergebnisse.

Die drei Algorithmen für die Partitionierung sollen die verschiedenen Ideen und Herangehensweise demonstrieren, die zur Lösung des Problems existieren. Es gibt noch viele andere Algorithmen, z.B. Verfahren die auf der Grundlage des Lernens arbeiten, um sinnvolle Cluster zu bilden. Des Weiteren wird an dieser Stelle es den Anwender überlassen, welcher Algorithmus für den Equilus verwendet werden soll.

Eine Weiterführung dieser Arbeit bestünde darin, weitere Messungen zu entwickeln in der eine bewusste Bewegung ausgeführt werden soll. Somit wäre ein Beugetest nach links, rechts und nach hinten vorstellbar.

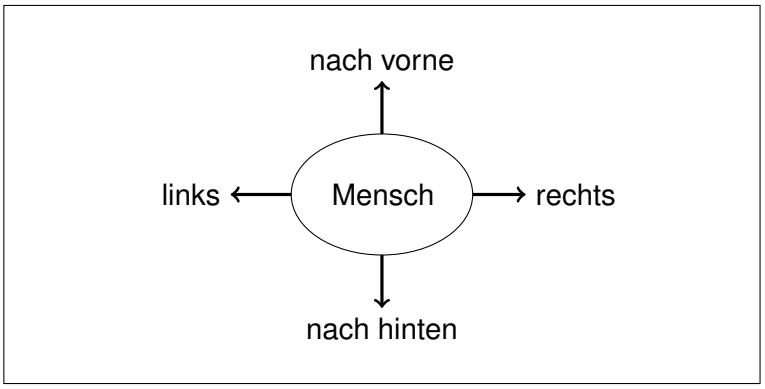


Abbildung 9.1: verschiedene Erweiterungen für den Beugetest

Der Kernpunkt in dieser Arbeit liegt darin, dass man eine Grundhaltung einnehmen muss und dann aus dieser Haltung eine Bewegung durchzuführen hat, um im Anschluss eine bestimmte Position zu halten. Darüber hinaus ist eine Rückkehr zur Ausgangsposition möglich, was mit Hilfe der Clusteranalyse verarbeitet werden kann.

Aus diesem Grund ist man in der Lage die verschiedenen Bewegungen, z.B. nach vorne und nach rechts zu kombinieren. Ein solcher Test ist nun mit dem Hilfsmittel der Clusteranalyse analysierbar und interpretierbar.

Literaturverzeichnis

[BEF84] James C. Bezdek, Robert Ehrlich, and William Full. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2):191–203, 1984.

[BPW10] Johann Bacher, Andreas Pöge, and Knut Wenzig. *Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren*. Oldenbourg Verlag, 2010.

[Cep15] Cepheiden. k-Means-Algorithmus. `https://de.wikipedia.org/wiki/K-Means-Algorithmus`, März 2015.

[EKSX96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[GM98] Udo Grimmer and Hans-Joachim Mucha. Datensegmentierung mittels clusteranalyse. In *Data Mining*, pages 109–141. Springer, 1998.

[Grä15] Horst Gräbner. DBSCAN. `https://de.wikipedia.org/wiki/DBSCAN`, August 2015.

[HK13] Frank Höppner and Frank Klawonn. *Fuzzy-Clusteranalyse: Verfahren für die Bilderkennung, Klassifizierung und Datenanalyse*. Springer-Verlag, 2013.

[JKS+13] Karin K. Januzaj, Peer Kröger, Jörg Sander, Matthias Schubert, and Arthur Zimek. Knowledge Discovery in Databases. 2013.

[Mic09] Stasius Michael. Clusteranalyse Arten und Anwendungen. `https://www.matse.itc.rwth-aachen.de/dienste/public/show_document.php?id=7276`, Dezember 2009.

[RAC04] Ziad Rached, Fady Alajaji, and Lorne Campbell. The Kullback-Leibler divergence rate between Markov sources. *Information Theory, IEEE Transactions on*, 50(5):917–921, 2004.

[Reh14] Philipp H. Rehs. Verfahren zur Dimensionsreduktion. 2014.

[Tit00] Peter Tittmann. *Einführung in die Kombinatorik*. Springer, 2000.

[WKM11] Katja Windt, Mathias Knollmann, and Mirja Meyer. Anwendung von Data Mining Methoden zur Wissensgenerierung in der Logistik-Kritische Reflexion der Analysefähigkeit zur Termintreueverbesserung. *Wissensarbeit-Zwischen strengen Prozessen und kreativem Spielraum*, pages 223–249, 2011.

[YLL12] Miin-Shen Yang, Chien-Yo Lai, and Chih-Ying Lin. A robust EM clustering algorithm for Gaussian mixture models. *Pattern Recognition*, 45(11):3950–3961, 2012.

Erklärung

Hiermit erkläre ich, dass ich meine Arbeit selbstständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die Arbeit noch nicht anderweitig für Prüfungszwecke vorgelegt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Mittweida, 06. September 2015